Complex networks across fields: from climate variability to online dynamics

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

d o c t o r r e r u m n a t u r a l i u m
(Dr. rer. nat.)
im Fach Physik
Spezialisierung: Theoretische Physik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
Frederik Peter Wilhelm Wolf, MSc.

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter:

1. Prof. Dr. Dr. h.c. mult. Jürgen Kurths
2. Prof. Dr. Francisco Rodrigues
3. Prof. Dr. Holger Lange

Tag der mündlichen Prüfung: 07.05.2021

To the bravely working cells in my myocard.
Without you, I would not have had the joy in the meantime.

Abstract

Complex networks are powerful tools enabling the study of complex systems. In many fields, complex networks are used as a tool to gain an understanding of the dynamics of interacting entities such as neurons in a brain, humans on social media, or global weather systems. At the same time, new theoretical frameworks that extend the toolbox of Network Science promote the application of network tools in new research fields. In this thesis, we aim for both, advancing the theoretical framework of Network Science as well as applying complex networks in Climatology and Computational Social Science.

Real-world networks are not only characterized by their topology but often also by their embedding in a two-dimensional spherical surface representing planet Earth. For such networks, we propose a framework to extract information from the network geometry in the first part of the thesis. In particular, we quantify the heterogeneity of the spatial arrangement of edges in different networks to illustrate the broad applicability of our method.

In the second part of the thesis, we shift the focus to functional climate networks as an application of complex networks in Climatology. By comparing two state-of-the-art event synchrony measures, we shed light on the caveats of network inference. In two case studies analyzing synchronous rainfall related to the South American Monsoon System, we show the distinct advantages of two event synchrony measures. With the analysis framework set up, we investigate the temporal evolution of the East Asian Summer Monsoon and identify a double band structure of synchronous heavy rainfall. By examining the characteristics of this rainfall band, we find that it is not only coinciding with the onset and withdrawal of the Baiu frontal system but also related to two atmospheric drivers that have previously been considered to be independent. Moving on from real-world climate networks to networks based on idealized model output data, we prove that global sea surface temperature variability is intimately linked to the time mean position of the inter-tropical convergence zone.

The concept of networks is not limited to studying climate systems, but has broad applicabilities. In this spirit, we utilize Network Science to contribute to Computational Social Science in the third part. Nowadays, online platforms are heavily scrutinized but undoubtedly influential actors in public discourse. Half of the Earth's population uses social media, therefore contributing to the rising amount of accessible data. Simultaneously, new trends emerge, gain and lose popularity with accelerating rate. Our contribution to the analysis framework of such data is a novel matching scheme which allows for tracking the evolution of network communities. Analyzing hashtag co-occurrence networks utilizing the proposed framework enables studying and characterizing the ups and downs of online discussions. In the last chapter of the thesis, we perform a large-scale study examining individual behavior on the social media platform *Twitter*. Exploring the trajectory of $600,000$ users, we find empirical evidence for an acceleration of individual user interactions. We believe that our observation of a shrinking individual content horizon due to the quantitative change in tweeting behavior can facilitate multiple problematic developments such as the increasing spread of false information on social media.

Zusammenfassung

Komplexe Netzwerke sind mächtige Werkzeuge, die die Untersuchung komplexer Systeme unterstützen. In vielen Bereichen werden komplexe Netzwerke eingesetzt, um die Dynamik interagierender Entitäten wie Neuronen, Menschen oder sogar Wettersysteme zu verstehen. Darüber hinaus erweitern sich die Anwendungsbereiche mit der stetigen Entwicklung neuer theoretischer Ansätze. In dieser Arbeit wollen wir sowohl den theoretischen Rahmen der Netzwerkwissenschaften weiterentwickeln als auch komplexe Netzwerke in der Klimatologie und der computergestützten Sozialwissenschaft anwenden.

Reale Netzwerke zeichnen sich nicht nur durch ihre Topologie aus, sondern oft auch durch ihre Einbettung auf eine zweidimensionale Kugel, den Planet Erde. Für solche Netzwerke führen wir einen theoretischen Ansatz ein, um Informationen aus der Netzwerkgeometrie solcher eingebetteter Netzwerke zu extrahieren. Anschließend quantifizieren wir die Heterogenität der Anordnung von Verbindungen in Netzwerken, um die breite Nützlichkeit unserer Methode zu veranschaulichen.

Im zweiten Teil der Arbeit konzentrieren wir uns insbesondere auf funktionale Klimanetzwerke, die eine Anwendung komplexer Netzwerke in der Klimatologie darstellen. Durch den umfassenden Vergleich zweier moderner Ereignissynchronisationsmethoden veranschaulichen wir die Herausforderungen der Netzwerkinferenz basierend auf Klimadaten. In zwei entsprechenden Fallstudien zur Analyse synchroner Niederschlagsereignisse im Zusammenhang mit dem südamerikanischen Monsunsystem zeigen wir die jeweiligen Vorteile der beiden Ereignissynchronisationsmethoden auf. Anschließend untersuchen wir die zeitliche Entwicklung des ostasiatischen Sommermonsuns mittels einer der beschriebenen Synchronisationsmethoden. Mithilfe funktionaler Klimanetzwerke identifizieren wir eine Doppelbandstruktur synchroner Starkniederschläge. Wir stellen fest, dass die Bildung und das Verschwinden nicht nur mit dem Einsetzen und Zurückziehen des Baiu-Frontensystems zusammenfällt, sondern auch mit zwei synoptischen Faktoren zusammenhängt, die zuvor als unabhängig angesehen wurden. Durch die nachfolgende Netzwerknanalyse von Daten, die aus idealisierten Klimamodellen gewonnen werden, zeigen wir, dass die globale Variabilität der Meeresoberflächentemperatur eng mit der Position der Innertropischen Konvergenzzone im zeitlichen Mittel korreliert.

Schließlich nutzen wir im dritten Teil Methoden der Netzwerkwissenschaft, um einen Beitrag zur computergestützten Sozialwissenschaft zu leisten. Heutzutage agieren soziale Medien als wichtige Akteure in öffentlichen Diskursen. Dazu ist die Hälfte der Weltbevölkerung in sozialen Medien aktiv und trägt somit zur steigenden Datenmenge bei, auf die zugegriffen werden kann. Gleichzeitig entstehen neue Trends, die stetig an Popularität gewinnen und verlieren. Um zur Verbesserung des theoretischen Analyserahmens beizutragen, führen wir zunächst einen neuartigen Algorithmus ein, mit dem die Entwicklung von Netzwerk-Communities verfolgt werden kann. Das Analysieren von Netzwerken, die wir durch die gleichzeitige Verwendung von Hashtags erstellen, erlaubt uns Beliebtheit in sozialen Medien zu untersuchen und zu charakterisieren. Im letzten Kapitel der Arbeit führen wir eine groß angelegte Studie einzelner Benutzer auf der Social-Media-Plattform *Twitter* durch. Durch das Verfolgen von 600.000

Benutzern finden wir empirische Belege für eine Beschleunigung individueller Benutzerinteraktionen. Wir sind der Meinung, dass unsere Beobachtung eines schrumpfenden individuellen Inhaltshorizonts aufgrund der quantitativen Änderung des Tweeting-Verhaltens mehrere problematische Entwicklungen wie die zunehmende Verbreitung falscher Informationen in sozialen Medien ermöglichen kann.

List of publications

This dissertation is partly based on the following publications. The identifiers P_1 - P_7 are cited in the text to highlight passages that are connected to these papers. The dissertation partly uses passages from the papers that were written by the author of this thesis. All passages provided by my co-authors were thoroughly rewritten.

P1 **Wolf, F.**, Kirsch, C., Donner, R. V.(2019). Edge directionality properties in complex spherical networks. Physical Review E 99.1., 012301

P2 **Wolf, F.**, Bauer, J., Boers, N., Donner, R. V. (2020). Event synchrony measures for functional climate network analysis: A case study on South American rainfall dynamics. Chaos 30.3., 033102

P3 **Wolf, F.**, and Donner, R. V. (submitted to European Journal of Physics - Special Topics). Spatial organization of functional climate network characteristics describing event synchrony of heavy precipitation

P4 **Wolf, F.**, Ozturk, U., Cheung, K., Donner, R. V. (2020). Spatiotemporal synchronization patterns of heavy rainfall events during the East Asian Baiu season. Earth Syst. Dynam. Discuss., in review

P5 **Wolf, F.**, Voigt, A., Donner, R. V. (2020). A climate network perspective of the intertropical convergence zone, Earth Syst. Dynam. Discuss., in review

P6 Lorenz-Spreen, P., **Wolf, F.**, Braun, J., Ghoshal, G., Djurdjevac Conrad, N., Hövel, P. (2018). Tracking online topics over time: understanding dynamic hashtag communities, Computational Social Networks 5.9

P7 **Wolf, F.**, Lorenz-Spreen, P., Lehmann, S. (in preparation). Generations of Twitter users reveal increasing engagement and shrinking content horizons

Berlin, May 21, 2021

Acknowledgements

First and foremost, I want to deeply thank Professor Jürgen Kurths for offering me the opportunity to pursue my Ph.D. in the exciting environment of the Potsdam Institute for Climate Impact Research and the Humboldt-University of Berlin. Likewise, I want to highlight the role of Professor Reik Donner, who has been my day-to-day supervisor. Without his questions, hints, and ideas I would not have successfully finished this endeavor.

In addition to the great institutional setting and the superb supervision, I want to thank my collaborators in the various projects. In this scope, I need to mention Dr. Philipp Gert Josef Lorenz-Spreen, who has been an inspiration during all three years. His endless optimism (in combination with some short periods of devastation ;)) has been a great motivation.

My daily work has also been immensely supported by my colleagues at PIK, in Brazil, and in the Portuguese class. Max, Jaqueline, Thomas, Nico, Fabi, Catrin K., Catrin C., Hauke, and Robertschi have been around me and helped by in various circumstances. Especially Max, Jaqueline, and Catrin K. helped me solving programming issues and frequently provided assistance when needed. To this end, I also want to thank Basti, Antje, Anton, and Pia who helped proofreading this manuscript.

Finally, I want to thank my whole family, my chosen family (Basti, Kita, Laura, Kati, Dän), and my friends who have backed me up in all situations. I credit my great spirit despite my health issues to you!

This work has been financially supported by the IRTG 1740/TRP 2011/50151-0, funded by the DFG/FAPESP.

Contents

List of frequently used mathematical symbols

Symbol	Meaning
N	Number of nodes in a network
E	Number of edges in a network
$\boldsymbol{A}(^t)$	(Temporal) adjacency matrix
a_{ij}	Element of adjacency matrix
$\boldsymbol{W}$	Weight matrix
ρ	Link density
$l(i,j)$	Shortest path length between node i and node j
D	Network diameter
T	Network transitivity
Q	Network modularity
L	Laplacian matrix
O	Average description length
$\boldsymbol{S}^{\square}$	Similarity matrix of differing origin
s_{ij}	Similarity matrix element
k_i	Degree of node i
Δk_i	Divergence of node i
v_i	Closeness of node i
b_i	Betweenness of node i
c_i	Local clustering coefficient of node i
d_i	Average link distance of node i
$d(e)$	Link distance of edge e
r_i	Local anisotropy of node i
p_i	Pairing coefficient of timeseries at node i
$\vec{r}_i$	Mean edge direction of node i

τ_{lm}^{ij}	Local dynamic coincidence interval of the Event Synchronization
τ	Time lag
ΔT	Global coincidence interval of the Event Coincidence Analysis
s_i	Number of events in time series i
$g(\cdot, \cdot)$	Pearson correlation coefficient
ϕ	Latitude
λ	Longitude
$\Theta(\cdot)$	Heaviside step function
$W\left(\{A_{t-n}, ..., A_{t-1}\}, B_t\right)$	Memory weight of community B and historical instances of A
$S_i(t)$	Size of community i at t

Part I.

Introduction, Theory and Methods

In this first part of the thesis, we introduce network science which offers powerful tools to study various systems of coupled entities. Specifically, we first outline its importance for the two main fields of applications in this thesis, Climatology and Computational Social Science in chapter 1. In the subsequent theoretical foundations (chapter 2), the basic concepts of Network Science are laid out. Here, the whole range from node-based network measures to global network characteristics is covered. This includes the discussion of mesoscale structures in networks, so-called network communities, before the multiple ways to construct networks in different application fields are reviewed. In the final chapter of the first part (chapter 3) a geometric approach to study spatially embedded networks is introduced. The presented framework is applicable in many different contexts and, thus, bridges from the theoretical foundations to the subsequent part where Network Science is used to study the Earth's climate system.

Chapter 1.

Introduction

In various situations, sensors, cameras, satellites, humans, or communication devices collect data [1]. In modern times, this data is utilized to analyze human behavior, predict stock market prices, understand the formation of climate feedbacks, plan responses to natural hazards, or improve all different kinds of services [2–6]. This development has led to the digitalization of many aspects of our modern society, including scientific research [7]. One of the most pressing challenges of research in the last two decades has been to make use of the increasing amount of available data [1, 8]. In this context, data science has emerged and has led to the development of powerful tools to process and abstract data into a form that is useful for contemporary computational methods [4, 9]. Located at the interface of computer science, physics, and statistics, data science makes use of many different approaches to generate insights from collected data. Having emerged as one branch of graph theory for analyzing relational data, nowadays, Network Science supports scientists from various disciplines in their endeavor to distill information from data [10].

As network scientists, researchers have contributed and still contribute to a multitude of different improvements. On the one side, theoretical work provides a growing number of tools which facilitate the detailed study of various systems [11–13]. On the other side, scientists from diverse backgrounds apply such tools to analyze systems ranging from coupled oscillators to transportation networks in cities [14, 15]. In this spirit, we attempt to promote data-driven investigations of dynamic processes using complex networks in this thesis. In particular, we propose a set of novel tools to study a specific class of networks as well as illustrate the broad applicability of complex networks by conducting studies in Climatology and Computational Social Science.

1.1. Complex networks in Climatology

The Earth's climate is a highly complex system [16]. Driven by solar insolation, an uncountable number of entities interact on different spatial and temporal scales [17]. Since prehistoric times, humans have tried to understand and forecast future weather and climate, relying, among others, on the seasonal cycle [18]. Anomalies and extreme events, such as heavy rainfall, droughts and temperature highs have influenced the evolution of humankind leading to conflicts and migration in the past and present [19, 20]. Nowadays, many components of the Earth's climate system have been identified and are well described in the literature but especially the accurate

prediction of climate change and corresponding shifts in the frequency and extent of extreme weather events are still challenges that Earth system scientists focus on [21–24]. This research is supported by a rising amount of data that is collected via satellites at fine spatial and temporal resolution, enabling climate scientists to progressively utilize data-driven approaches [25].

Almost two decades ago, network approaches have entered the field of Climatology [26, 27]. So-called climate networks act as functional representations of the underlying climate system. They are often constructed via a non-linear correlation measure using observations from various locations which are commonly distributed on a regular spherical grid [28–30]. Incorporating non-linearity enables the corresponding network representations to complement traditional linear methods and has led to multiple discoveries. Notable examples include a newly developed network-based index that allows for discriminating between different flavors of the El Niño–Southern Oscillation (ENSO) [31], a network-based prediction scheme that has improved the Indian monsoon forecast [32], and the identification of Rossby waves as a source for global teleconnections of extreme rainfall [24].

Along with the exploration of the Earth's climate system using data-driven approaches, new methods for constructing and analyzing networks have been proposed and utilized. Event Synchronization, which is one of the most frequently used methods to correlate climate event time series, was initially developed for measuring the synchrony of EEG spikes. The utilization of Event Synchronization in Climatology is an example of interdisciplinary method transfer in the context of complex networks [33, 34]. In the second part of the thesis, we do not only apply Event Synchronization to analyze precipitation time series, but we also investigate possible biases of synchrony measures in general. Furthermore, we outline two studies where we examine synchronous rainfall related to the East Asian Summer Monsoon and classify model outputs from different classes of climate models utilizing zonal mean network measures of a surface temperature-based network.

1.2. Complex networks in Computational Social Science

With the popularity of online platforms such as *Facebook*, *Twitter*, or *Instagram*, the organization of societal communication has transformed dramatically [35, 36]. Motivated by the increasingly important role of these communication channels, the emerging data source for quantifying human behavior, and the rising computational power available, computer scientists, physicists, or more generally, data scientists, have entered the Social Sciences [37]. Their collaboration with researchers from the Social Sciences has yielded tools to analyze data that was generated by the interaction of humans [38, 39]. These interactions can be of different kind, ranging from measuring contacts based on the spatial distance of people at conferences to friendship relations on some social network. In addition, the described interplay of people, or entities is most often highly dynamic and thus, changes its characteristics over time [40, 41].

One side effect of the associated acceleration of information flow is the increased spread of misinformation [42, 43]. Besides, complex effects of social network topologies can even lead to undemocratic decisions [44]. Therefore, it is not only important to examine the new forms of interaction and to understand corresponding implications for the formation of opinions. Also investigating the consequences of changes or interventions that get implemented on online platforms are pressing topics of Computational Social Science.

Due to the multitude of different opportunities to interact in online environments, networks can be defined in various ways from corresponding data sets. In the last part of the thesis, we employ two distinct strategies to infer networks from social interaction. In the first chapter of the third part of this thesis, we propose a method to construct content networks from the co-occurrence of hashtags and study the associated dynamic changes of the network structure. In the second chapter, we investigate interaction patterns of individual users on Twitter in a longitudinal study and illustrate how online platforms and their users' behavior have changed over time.

1.3. Organization of the thesis

To subdivide between the theory and different application contexts, we have split the thesis into three parts.

In the first part, the introduction (chapter 1) is followed by the theoretical foundations of the conducted analyses (chapter 2). These sections serve as the toolbox which is utilized in the second and third part of the thesis. We start by introducing networks as abstract objects and then we present an overview of the basic concepts of Network Science. This includes reviewing commonly applied global and local network measures which are used in the scope of the thesis. In addition, we discuss state-of-the-art approaches to group nodes into distinct clusters and present different strategies to analyze temporal changes of network topology. To connect these theoretical concepts with the broad scope of applications, we continue illustrating the construction of functional climate networks by reviewing the most frequently used strategies to construct networks from climate time series. In Climatology, not only network construction comes with different obstacles, also the analysis of networks where nodes are embedded in physical space requires specialized tools. Accordingly, spatially embedded networks and associated network measures are discussed. Along with the introduction of the common concepts, we examine differently sourced biases and the corresponding correction schemes. The theoretical foundations are closed by a short comment on the construction of networks in Computational Social Science. As there are multiple different ways to design networks from data of human (online) interaction, we exemplarily highlight two distinct approaches. From chapter 3 onward, we present our own original results. We initially extend the toolbox of Network Science by introducing geometric network measures suited for studying networks embedded in spherical geometry. We therein propose a comprehensive set of edge directionality measures, which can be defined in spatially embedded networks. By studying edge

directionality in complex climate networks as well as in an air transportation network and in the world trade network of 2009, the broad applicability of geometric network measures is illustrated and already indicates the potential of network approaches. Therefore, we consider this mainly methodological work as a bridging chapter connecting the theoretical foundations with the applications presented in part two and part three of the thesis.

In the second part of the thesis, we make use of these theoretical foundations and discuss applications of climate networks. To highlight the advantages of two distinct event synchrony measures that are employed to construct climate networks in this thesis, two case studies are conducted in chapter 4. Here, we initially focus on differently caused shortcomings and discuss a suited correction scheme, before illustrating the respective capabilities to track cascades of heavy rainfall related to the South American Summer Monsoon. Subsequently, Event Coincidence Analysis, one of the previously studied event synchrony measures, is used to identify patterns of synchronous rainfall related to the East Asian Monsoon System in chapter 5. In particular, we study the Baiu front which distributes heavy rainfall in East Asia and show that the emergence of a double band structure of synchronous heavy rainfall is closely related to the onset of Baiu related heavy rainfall south of the Japanese archipelago. As a second application of climate networks, an ensemble of model outputs is examined in chapter 6. Specifically, we investigate idealized aquaplanet simulations from different comprehensive global circulation models. By analyzing the zonal mean network measures of the various corresponding network representations, we find that the surface air temperature correlation patterns are related to the time-mean position of the inter-tropical convergence zone (ITCZ). By clustering the models according to their zonal mean network measures, we identify groups of climate models which exhibit distinct mean ITCZ positions. Therein, the related network charcteristics of different groups are discussed.

In the third part, we change the area of application and study networks in Computational Social Science. In the first study, we construct networks from co-occurrences of hashtags in chapter 7 showing that networks derived from posts in an online fashion blog exhibit strong modularity and are, thus, ideally suited to track the popularity of different topics. In this scope, we introduce a comprehensive matching scheme to connect the subsequent temporal layers of the hashtag co-occurrence network. We additionally illustrate how the incorporation of memory in the matching framework enables us to successfully track the evolution of network communities. In chapter 8, we conduct a study based on a very large subset of interactions on the social media platform *Twitter*. By analyzing posts from $600,000$ users that have been active throughout the last eight years, we show how social platforms have changed over time. Furthermore, we confirm that users individually behave differently depending on the provided platform environment. Observing the growing responsivity and, thus, the increasing burstiness of user interaction, we prove social acceleration on an individual level.

Finally, in chapter 9, we summarize the main conclusions derived in chapters 3-8. We do not only highlight the variety of application fields of complex networks and

state our main results, but also elaborate on future directions of research in the different disciplines.

Chapter 2.

Theoretical foundations

Complex systems are omnipresent. Ranging from microscopic feedback loops in the human brain to the interaction within galaxy clusters, complex systems can be found almost everywhere and can be identified at all temporal and spatial scales [45–47]. Characterizing such systems, understanding their dynamics, and even forecasting future developments challenge humanity since millennia [48]. While theoretical physics and mathematics have contributed for centuries to grasp the nature around us [49], computational methods have just recently become relevant. Among the multiple methods to study complex systems, complex networks are one tool enabeling the analysis of relational data. Originating from graph theory which has been a branch of mathematics since the 18th-century [50], Network Science is presently applied in a multitude of disciplines [10].

Networks are often abstract representations of a set of interactions. Thus, it is not always only a network itself that is being studied. Also, network construction is sometimes challenging. In this thesis, networks are obtained in many different ways based on a variety of interactions and various sorts of data. Therefore, we first introduce networks as abstract objects and introduce common measures to characterize the interacting entities before we outline strategies to construct networks from different data sets.

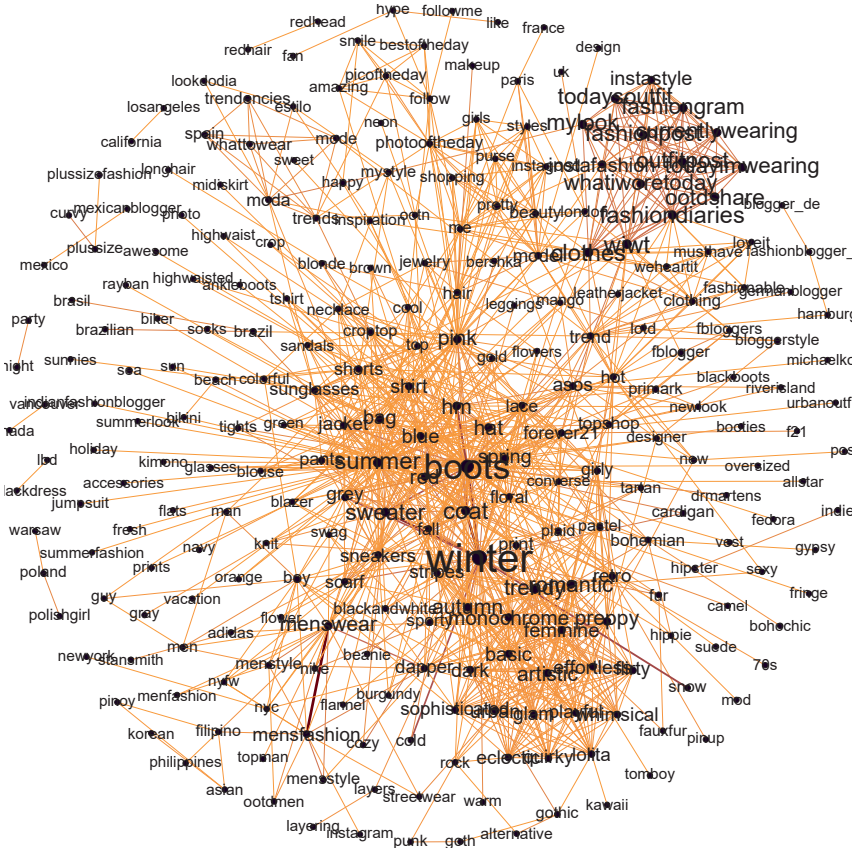

Figure 2.1.: Example of a weighted and undirected empirical real-world network. The network shows co-occurences of hashtags on the fashion blog `lookbook.nu` (for details, see chapter 7). Node label sizes scale with the node degree. Edge color indicates edge weight. Figure designed with gephi [51] using the Fruchterman-Reingold representation.

2.1. Network science

2.1.1. Adjacency matrix

A network consists of N nodes which are connected by E edges which can be directed, undirected, weighted, and unweighted. In Fig. 2.1, we show a weighted and undirected network obtained from a data set which we will study later in this thesis. The topology of such networks is most commonly encoded in the so-called adjacency matrix $\boldsymbol{A}$. In the case of an unweighted and undirected graph, the elements of the adjacency matrix take the value of $a_{ij} = a_{ji} = 1$ if node i and j are connected and $a_{ij} = a_{ji} = 0$ if they are not connected [52]. Such a mathematical representation offers the opportunity to compute a multitude of network measures to analyze a network's properties.

To illustrate the usability of the adjacency matrix, we note that we can infer topological path length (graph distance between two nodes) directly from basic

calculations. For a pair of nodes i and j, we first check whether the matrix element a_{ij} equals 1. In this case, we can stop the calculations and conclude that there is a direct connection between node i and node j. If not, we compute $\boldsymbol{A}^2$. An element $a^2_{i,j}$ of this matrix takes value $a^2_{i,j} = 1$ if there exists a 2-step path between node i and node j. If not, $a^2_{i,j} = 0$ and we proceed by computing $\boldsymbol{A}^3$. Following this rationale, we can iteratively determine the path length between each pair of nodes.

Weights of edges in a network are often motivated by systematic differences between edges and substantial discrepancies in the importance of edges. To account for such weighted edges we can replace the adjacency matrix by the weight matrix $\boldsymbol{W}$ with elements w_{ij} representing the weights of the edges between nodes i and j [53]. In the following, we will omit the discussion of possibly weighted edges, as most measures can be extended in a straightforward manner to weighted networks. In subsequent parts where we discuss weighted networks, we specify the utilized measures and their particular definition.

Some networks represent unidirectional relations which are not well captured by bidirectional edges. In this case, the elements a_{ij} and a_{ji} of the adjacency matrix (or the corresponding values in the weight matrix) are not equal (since an unidirectional edge pointing from node j to node i is completely reflected by $a_{ij} = 1$) [54]. In the following, we usually introduce the network measures for undirected networks, as most of the studied networks in this work are undirected and the majority of the network measures can be easily adapted to the study of directed networks.

2.1.2. Global network measures

In a network with N nodes, the maximal number of edges is given by $E_{\max} = N \cdot (N-1)$. In networks, where a node can be connected with itself via a so-called self-loop (normally modeling some kind of self-feedback of some entity), the total number of edges can reach up to $E_{\max} = N^2$. In most real-world networks as well as in functional networks, where nodes and edges model the interaction in phyical systems, only a few of the possible edges are present. To quantify the number of edges present in a network we commonly calculate the *link density* ρ by [30]

$$\rho = \frac{E}{E_{max}} = \frac{\sum_{ij} a_{ij}}{N(N-1)}. \tag{2.1}$$

If all edges are present in a (sub-)network or graph, we call this object *complete*.

To o obtain a rough knowledge of the global arrangement and interconnectivity of a network, we can compute the *average shortest path length* l by [10]

$$\langle l(i,j) \rangle = \frac{1}{N(N-1)} \sum_{i \neq j}^{N} l(i,j). \tag{2.2}$$

Here, l measures the shortest path length between node i and j which denotes the minimal number of edges an imaginary walker on the network would have to traverse to reach node j starting at node i.

The maximum of all shortest path lengths in a network is commonly known as the network *diameter* D [10]

$$D = \max_{i \neq j}^{N} l(i, j). \tag{2.3}$$

To quantify the existence of transitive links within a network, we can calculate the *network transitivity* T. The network transitivity is the ratio between the number of topological triangles in a network and the number of connected triples in the network. We can compute the transitivity utilizing the adjacency matrix by [10]

$$T = \frac{\sum_{i,j,k=1}^{N} a_{ij} a_{ik} a_{jk}}{\sum_{i,j,k=1; j \neq k}^{N} a_{ij} a_{ik}}. \tag{2.4}$$

2.1.3. Node-wise network measures

A variety of node-wise defined quantities complement the previously introduced global network measures.

To investigate the role and importance of nodes in a network, various centrality measures have been proposed which often account for very specific properties. The *node degree* is the most basic of all centrality measures and simply equals the number of adjacent edges to a single node. This can mathematically be achieved by a summation over the columns of the adjacency matrix. In an undirected, unweighted network the *degree* k of node i reads [10]

$$k_i = \sum_{j=1}^{N} a_{ij}. \tag{2.5}$$

In directed networks, we can differentiate between in- and outgoing edges for each node. Therefore, we define the *in-degree* as

$$k_i^{in} = \sum_{j=1}^{N} a_{ij}. \tag{2.6}$$

Accordingly, we define the *out-degree* as

$$k_i^{out} = \sum_{j=1}^{N} a_{ji}. \tag{2.7}$$

In this manner, many network measures can be extended to directed networks.

A measure that is only defined for directed networks is the *network divergence* Δk. We compute the network divergence by considering the difference between in- and out-degree [55]

$$\Delta k_i = k_i^{in} - k_i^{out} \tag{2.8}$$

and interpret the network divergence as an indication of sinks and sources in a network.

In addition to node-based quantifications of adjacent edges, we can also compute the interconnectivity of different subgroups of nodes N_m and N_n. This is commonly achieved by calculating the *total cross degree* [56]

$$k^{m,n} = \sum_{i \in N_m} \sum_{j \in N_n} a_{ij}. \tag{2.9}$$

In contrast to the previously introduced network measures, which are directly related to the degree of a node, the *closeness* v is associated with the topological distance between the nodes in the network. Specifically, we compute the closeness by [30]

$$v_i = \frac{N-1}{\sum_j l(i,j)}. \tag{2.10}$$

Accordingly, nodes with a large closeness have on average shorter paths to other nodes in the network than nodes with a small closeness.

Following a different rationale, we can measure the centrality of a node by computing how many shortest paths in a network traverse this node. The corresponding centrality measure *betweenness* b is defined by [30]

$$b_i = \sum_{j \neq i \neq k}^{N} \frac{\sum_{j \neq k}^{N} l(j,k|i)}{\sum_{j \neq k}^{N} l(j,k)} \tag{2.11}$$

where i, j and k serve as node labels and $l(j,k|i)$ denotes a shortest path between node j and k which traverses node i in between.

To quantify the connectedness of the neighborhood of a node, there has been introduced the *local clustering coefficient* c [10]

$$c_i = \frac{2e_i}{k_i(k_i-1)}. \tag{2.12}$$

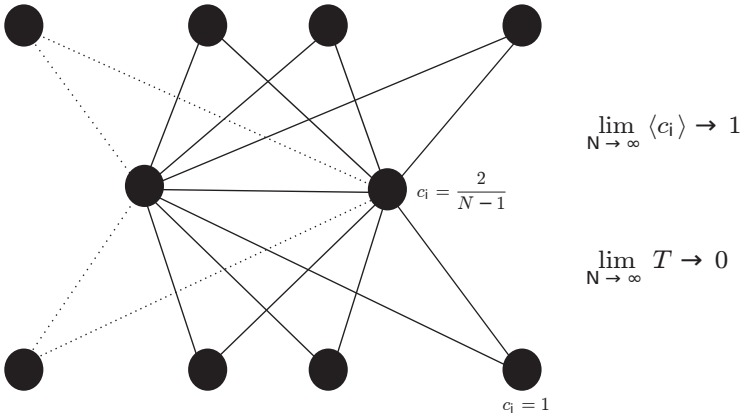

Figure 2.2.: Illustration of a network highlighting the difference between the global clustering coefficient and the network transitivity.

Here, e_i represents the number of edges connecting the topological neighbors of node i. The local clustering coefficient takes values between 0 (no neighboring nodes are connected) and 1 (neighbors form a complete subgraph). We can average this measure for all nodes in the network which is commonly known as the *global clustering coeffient*.

A measure that is often mistakenly assumed to be identical to the global clustering coefficient is the network transitivity (Eq. 2.4).

The similarity, as well as the difference between the transitivity and the clustering coefficient, can be directly inferred by considering an alternative form of Eq. 2.12 [30]

$$c_i = \frac{2e_i}{k_i(k_i-1)} = \frac{\sum_{j,k=1}^{N} a_{ij}a_{ik}a_{jk}}{k_i(k_i-1)}. \tag{2.13}$$

This definition directly points to the systematic difference between these two measures. Whereas the local clustering coefficient and, therefore, also the global clustering coefficient directly depends on the degree k of the nodes in the network, the transitivity relates to the overall connectivity pattern and does not explicitly depend on the degree k. We illustrate this by utilizing two examples: first, we consider a connected pair of nodes that are each the center of star-like network topologies and are connected to all nodes (see Fig. 2.2). In such a network, each node (beside the two nodes at the centers) has clustering coefficient 1 which leads to $\frac{1}{N}\sum_{i=1}^{N} c_i \to 1$ for large N. In turn, the transitivity in such a topology decreases due to the absence of transitive connections (see Fig. 2.2) [10]. Second, in real-world networks, it has been shown that the emergence of localized structures goes hand in hand with the simultaneous decrease of the global clustering coefficient as normally the degree of such nodes increases. In the same scenario, the transitivity has been shown to still increase due to the influence of new transitive links in the network [57].

After the introduction of globally (network diameter, average shortest path length,...) and locally (degree, local clustering coefficient,...) defined network measures, we next present network communities, which describe mesoscale structures in complex networks.

2.1.4. Communities

To analyze a network at mesoscale, the concept of network communities has been developed [10, 58, 59]. A community within a network describes a set of nodes that is highly interconnected and exhibits fewer connections to the rest of the network. Such communities are most commonly derived by unsupervised algorithms and can have many different attributes. Accordingly, the specific definition of a community differs among the algorithms. Whereas some methods allow for overlapping communities and do not attribute all nodes to some community, others classify each node as a member of one distinct community [60–62].

A first class of community detection algorithms uses the concept of modularity to obtain network communities [63]. Modularity is a measure that quantifies the difference between the edge density within a set of nodes and the density expected from a random graph [64]. Thus the modularity reads [10]

$$Q = \frac{1}{2E} \sum_{i,j} \left(a_{ij} - \frac{k_i k_j}{2E} \right) \delta_{X_i, X_j} \tag{2.14}$$

given the community affiliation X_i of node i and the probability of a link between the respective nodes in a random graph model $\frac{k_i k_j}{2E}$. Modularity maximization leads to a complete partitioning of the underlying network. In addition, the modularity often exhibits a plateau around its maximum and, therefore, shows results of partitioning that are rather unstable to small changes of the network topology. Moreover, the modularity is characterized by a resolution limit for small communities. Therefore, more sophisticated algorithms have been developed.

A second class of community detection algorithms uses random walkers to identify highly interconnected sets of nodes [40, 65–67]. This approach is motivated by the observation that random walkers might get stuck for some time in densely connected subgraphs. To describe random walks on a network, methods often rely on the Laplacian matrix which is defined by [10]

$$L_{ij} = \begin{cases} k_i, & \text{if } i = j, \\ -a_{ij}, & \text{otherwise,} \end{cases} \tag{2.15}$$

for an undirected network. Note that the sum of each row (and column) of the Laplacian is zero which directly corresponds to the application of the Laplacian operator to describe diffusion processes. To directly incorporate the transition

probabilities into the Laplacian, we can normalize each row by the corresponding degree k_i.

One popular random-walker-based method is the Infomap algorithm [65]. Starting from trajectories of random walkers on the network, the algorithm relies on the minimization of the corresponding description length. The map equation [65]

$$O(M) = q_{\curvearrowright} H(q) + \sum_{i=1}^{n} p^i_{\circlearrowright} H\left(p^i_{\circlearrowright}\right) \tag{2.16}$$

describes the average code length which is needed to specify a single step of a random walker trajectory on a network with partitioning M. The first part accounts for the jumps between the n communities and the second for the jumps and the respective code length within each community. $H(\cdot)$ describes the entropy of the community names or the within-community movements, respectively. The application of the Infomap algorithm does not only lead to more robust communities but is also applicable for all kinds of networks.

In addition to these two classes of community detection algorithms, there exist multiple other systematically different strategies. In particular, the statistical inference has been successful in contributing new insights via community detection [68]. As we are not specifically interested in presenting a comprehensive overview of community detection algorithms in this thesis, we refer to the corresponding literature for further details [68, 69]. In this thesis, we only identify communities using modularity maximization and the Infomap algorithm.

2.1.5. Temporal networks

As networks are sometimes abstract models of physical processes that change in time, we incorporate temporality in the analysis of networks [41, 70]. As an example, temporality is crucial to correctly describe information diffusion between agents in a contact network [71, 72]. Let us assume that nodes in such a network are persons which at some time meet each other and exchange the information which thereby diffuses among the network's nodes. In such a scenario, analyzing the static aggregated network of all contacts over a longer period can only be a poor approximation of the actual diffusion process. In this example, the temporal order of the meetings between three persons is a decisive factor whether it is possible to pass the information. Such considerations are important in many different applications of networks as abstract models of physical processes [72].

There are two basic strategies to integrate temporality in networks. First, we can attribute a specific time or period to every edge and analyze the stream of edges. This enables us to fully grasp the interactions between the agents which are modeled as nodes of the network. Alternatively, we can aggregate the interactions for some short period and analyze consecutive time windows. Although this method is not as exact as considering the exact temporal order of edges, it is sufficient to describe the temporal change in some applications. In chapter 7, we investigate the dynamically

changing popularity of topics in a social network. In this specific context, it is more the coarse time frame of days and weeks at which the popularity of topics changes. We, therefore, stick to this approach in our analysis in chapter 7.

Following the rationale of aggregating interactions in some short time-window, we can generalize all the above-mentioned network measures by attributing a time domain to all of them. We start by obtaining a temporal (time-window-wise) adjacency matrix $\boldsymbol{A}^t$ which encodes the topology of the network for the particular time window at time t. Accordingly, we subsequently compute the time-dependent degree of node i, k_i^t. In this way, we construct temporal layers of networks with certain properties that we can track over time to reveal the dynamic change of the studied network.

2.2. Construction of functional climate networks

As mentioned above, in some contexts, network construction is not straightforward. This holds especially for networks in Climatology where edges represent some functional connection between nodes which represent locations in physical space. In the following, we describe how we infer the existence of such edges by using the similarity of time series.

Functional climate networks, as they were introduced by Tsonis et al. [26, 27] and further developed by Donges et al. [28], are obtained based on climatic time series from some variable. Since then, there have been presented multiple studies employing and improving the introduced frameworks. An overview over some publications using functional climate networks based on event synchrony can be found in appendix B. The general rationale behind climate networks is illustrated in Fig. 2.3 [30].

We start by selecting variables such as (sea surface) temperature or precipitation in a region of interest with an appropriate temporal and spatial resolution. Throughout the thesis, a variety of data sets with diverse spatial and temporal resolution from different sources are utilized. More details on the data are given in the respective sections.

In a second step, we calculate the pairwise similarity between all time series from the different locations within the region of interest (see Fig. 2.3). Depending on the similarity measure, we need to pre-process the data to be able to apply the selected similarity measure. In this thesis, we estimate the similarity between time series using Event Synchronization [33], Event Coincidence Analysis [73], and Pearson correlation [74].

Event Coincidence Analysis and Event Synchronization quantify the similarity between event time series. Therefore, we either need to directly access data which has an event-like structure or transform non-event time series into event time series. An established approach for the latter is to define events by thresholding the values of each time series at some percentile [32, 75]. The methodological details of Event Synchronization and Event coincidence analysis are described in detail in the subsequent sections (section 2.2.1 and section 2.2.2).

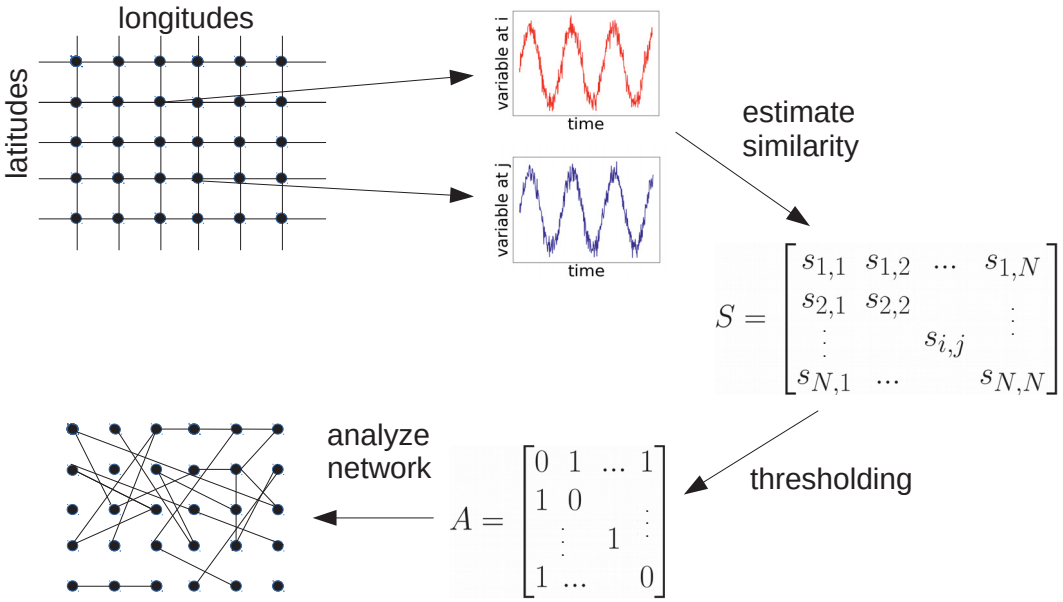

Figure 2.3.: Schematic illustration of climate network construction adapted from publication [30].

To successfully distill information from networks based on Pearson correlation, we follow the established procedure of first calculating the anomaly (deseasonalized) time series. This is commonly done by subtracting the annual mean climatology for each time series [28, 30]. Thereafter, we compute the correlation between the time series. We outline the details in a subsequent section (section 2.2.4).

Regardless of the similarity measure, we order the pairwise similarity scores in a quadratic similarity matrix $\boldsymbol{S}$ which we can directly interpret as a weighted adjacency matrix of a network in which each location of the respective time series represents a node (as illustrated in Fig. 2.3). Although we could start a network analysis directly at this point, this weighted adjacency is often transformed into a binary unweighted version. As the number of grid points is normally of order $N = 10^5 - 10^6$, keeping the weighted adjacency matrix does not only decrease computations speed but is sometimes not possible due to limited memory. Also, low similarity (arising from estimating the similarity of *all* pairs of time series) does most often have no physical meaning and might induce some biases in the network structure.

To binarize the matrix, two main approaches have been applied in the context of climate networks. On the one hand, we can use statistical testing to decide whether a similarity value is significant up to some confidence [5, 24]. On the other hand, we can threshold the similarity matrix at some value to only keep the strongest similarity and thereby control the resulting link density in the network [31, 75, 76].

In each chapter, we specify and motivate the choices for the methodological setup which we tailor differently for the distinct applications. There, we also present the details of the network analysis (the last step in the illustration in Fig. 2.3).

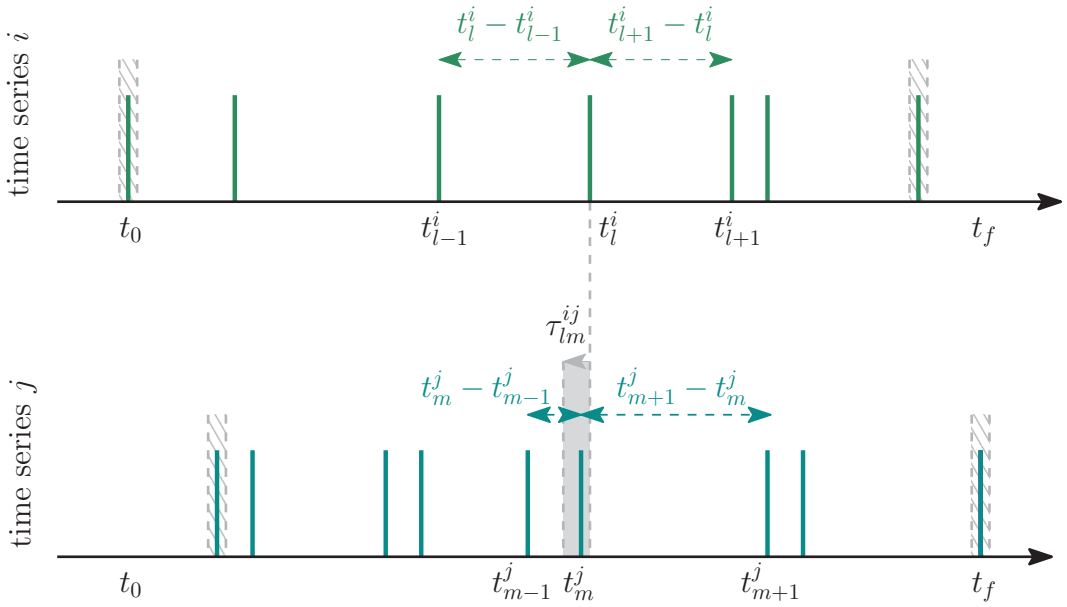

Figure 2.4.: Schematic illustration of synchrony estimation using ES. Depicted are two event time series i and j. Green lines indicate events. Striped areas are excluded in the computation for correct normalization. The dynamic, local coincidence window is shaded grey. Figure taken from publication P_2.

In the following sections, we introduce the rationales and definitions behind Event Synchronization, Event Coincidence Analysis, and Pearson correlation.

2.2.1. Event synchronization

Event Synchronization (ES), first introduced by Quiroga et al. [33], is a parameter-free method to quantify the synchrony of events in event time series. This method has frequently been utilized to construct networks from climate time series in the last decade. Pre-requisites for the application of ES are event time series for observations located at the different nodes in a network. In the following, we describe how we estimate the similarity between event time series and, thus, obtain an adjacency matrix. Here, we stick to the notation which has been presented in recent publications [77, 78].

Using ES, two events l and m in time series at node i and j at time t_l^i and t_m^j are considered to be *synchronized* if, and only if, they have occurred within the *local (dynamical) coincidence interval* (as shown in Fig. 2.4) [78]

$$\tau_{lm}^{ij} = \frac{1}{2} \min \left\{ t_{l+1}^i - t_l^i, t_l^i - t_{l-1}^i, t_{m+1}^j - t_m^j, t_m^j - t_{m-1}^j \right\}. \tag{2.17}$$

Thus, the synchrony condition reads

$$\sigma_{lm}^{ij} = \begin{cases} 1, & \text{if } 0 < t_l^i - t_m^j \leq \tau_{lm}^{ij}, \\ 0, & \text{otherwise,} \end{cases} \tag{2.18}$$

and can, in principle, be limited by an upper bound τ_{max} to avoid unrealistically large coincidence intervals.

As the local coincidence interval is calculated by subtracting temporal distances *between* events, we exclude the first and the last event of all time series [77].

To estimate the total number of synchronized events we compute

$$c(i|j) = \sum_{l=2}^{n_i-1} \sum_{m=2}^{n_j-1} J_{lm}^{ij}, \tag{2.19}$$

utilizing the indicator funtion [77]

$$J_{lm}^{ij} = \begin{cases} 1, & \text{if } \sigma_{lm}^{ij} = 1, \sigma_{m,l-1}^{ji} = 0 \text{ and } \sigma_{m+1,l}^{ji} = 0, \\ \frac{1}{2}, & \text{if either } t_l^i = t_m^j \\ & \text{or } \sigma_{lm}^{ij} = 1 \text{ and } (\sigma_{m,l-1}^{ji} = 1 \text{ or } \sigma_{m+1,l}^{ji} = 1), \\ 0, & \text{otherwise.} \end{cases} \tag{2.20}$$

Here, the definition of $c(i|j)$ implies that events in time series at node j precede events in time series at node i. We, therefore, have to use this version of the indicator function to prevent double counting. In addition, we set the total number of events as $l = 2, 3, ..., n_i - 1$ and $m = 2, 3, ..., n_j - 1$ due to the exclusion of the first and last time series.

We then calculate the event synchronzation strength by considering the numbers of synchronized events in both directions ($c\,(i|j)$ and $c\,(j|i)$)

$$s_{ij}^{ES} = \frac{c\,(i|j) + c\,(j|i)}{\sqrt{(n_i - 2)\,(n_j - 2)}}. \tag{2.21}$$

Note, that we exclude the first and the last event of each time series for a correct normalization. Finally, we use the symmetric matrix $\boldsymbol{S}^{ES} = \left(s_{ij}^{ES}\right)$ to construct a functional climate network. As there are several ways to transform $\boldsymbol{S}^{ES}$ into an adjacency matrix, we discuss different utilized options in the respective chapters.

2.2.2. Event Coincidence Analysis

Event Coincidence analysis (ECA) [73] quantifies the similarity between event time series in a similar fashion as ES. In contrast to ES where a dynamic, data-adaptive, coincidence interval is utilized, ECA follows the rationale of a static (globally defined) coincidence interval (set by a parameter), which can be adjusted to the desired time

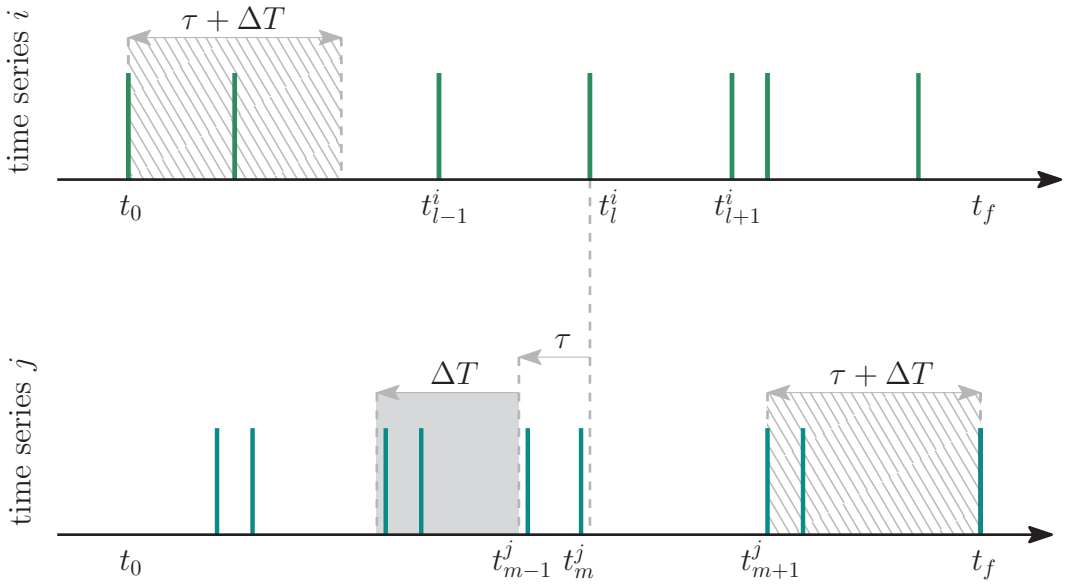

Figure 2.5.: Schematic illustration of lagged synchrony estimation using ECA. Shown are two event time series i and j. Green lines indicate events. Striped areas are excluded in the computation for correct normalization. The static, global coincidence window is shaded grey. Figure taken from publication P_2.

scale allowing for analytical examinations [73]. The basic idea of the method is illustrated in Fig. 2.5. Again, we stick to the notation which has been presented by Odenweller et al. [77].

Using ECA, two events at times t^i_l and t^j_m are considered to coincide if they occur within the global coincidence interval ΔT and, thus, fulfill the condition

$$0 < t^i_l - t^j_m < \Delta T. \tag{2.22}$$

The coincidence rates between the events from two event time series indicating the fraction of events in time series i that are preceded by at least one event in time series j are computed as [77]

$$r\left(i|j;\Delta T\right) = \frac{1}{n_j - n'_j} \sum_{m=1}^{n_j - n'_j} \Theta\left\{\sum_{l=1}^{n_i} 1_{[0,\Delta T]}\left(t^i_l - t^j_m\right)\right\} \tag{2.23}$$

with the indicator function

$$1_I\left(x\right) = \begin{cases} 1, & \text{if } x \in I, \\ 0, & \text{otherwise,} \end{cases} \tag{2.24}$$

and the Heaviside step function $\Theta(\cdot)$. Whereas the Heavyside step function prevents the double counting of events, removing the number of events within the time interval between $t_f - \Delta T$ and t_f allows for correct normalization. We, therefore, subtract

$$n'_j = \sum_{m=1}^{n_j} 1_{[t_f - \Delta T, t_f]} \left(t_l^i\right) \tag{2.25}$$

from the number of events in time series j.

At this point, we can again highlight the bidirectionality of undirected connections by calculating the mean of the event coincidence rates

$$s_{ij}^{ECA,mean} = \frac{r\left(i|j;\Delta T\right) + r\left(j|i;\Delta T\right)}{2} \tag{2.26}$$

for obtaining a symmetric similarity matrix $\boldsymbol{S}^{ECA}$.

As an additional feature, which we utilize in chapter 4 and chapter 5, we can also consider the the directional coincidence rates $r\left(i|j;\Delta T\right)$ and $r\left(j|i;\Delta T\right)$ to emphasize a strong unidirectional association between the event time series. In this case, we define elements of the similarity matrix $\boldsymbol{S}^{ECA}$ by computing the maximum of the two pairwise event coincidence rates

$$s_{ij}^{ECA,max} = \max\left(r\left(i|j;\Delta T\right), r\left(j|i;\Delta T\right)\right). \tag{2.27}$$

Note that examining the directionality of connections can, in the same manner, be computed using Event Synchronization.

In this thesis, we omit the discussion of possible differences between the trigger and precursor coincidence rates [73, 77] which differ regarding their normalization in Eq. 2.23. The stated definition refers to the trigger coincidence rate where events in time series i trigger events in time series j. Our computations have led to (visually) identical results using the precursor coincidence rate in the studied cases.

Finally, we employ the similarity matrix $\boldsymbol{S}^{ECA} = \left(s_{ij}^{ECA}\right)$ to obtain an adjacency matrix for the construction of functional climate networks.

2.2.3. Time Delayed Versions of ES and ECA

In the above definitions, both methods feature temporal distances ≥ 0 between two coinciding events. Whereas the introduction of a second parameter τ taking care of lagged coincidence is commonly discussed in the context of ECA, a similar parameter can be easily implemented in the ES serving as a lower limit of the dynamic coincidence interval. However, in the present thesis, we will only apply the lagged version ECA in chapter 4 while we utilize the ES only in the stated version. Regarding the previously introduced definitions, a time-lagged version of ES and ECA can be realized by a shift of one event sequence by the constant value τ in the equations of the two previous subsections [77].

2.2.4. Pearson correlation

Computing the linear Pearson correlation coefficient is an established way to estimate correlation between two samples. In this thesis, we utilize the Pearson correlation coefficient to access the similarity between to equally sampled functions (of time or latitude). For two functions x and y we define the Pearson correlation coefficient as [74]

$$g(x,y) = \frac{\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \hat{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \hat{y})^2)}} \tag{2.28}$$

Here, $\hat{x}$ and $\hat{y}$ denote the time-mean or latitudinal-mean of x and y. For constructing a climate network, we consider the (deseasonalized) time series measured at the grid points and compute the Pearson correlation coefficient for each pair (i,j). Due to the symmetry $g(x,y) = g(y,x)$ we obtain the symmetric similarity matrix $\boldsymbol{S}^{\text{Pearson}}$ with entries

$$s_{ij}^{P} = g(x_i, x_j) \tag{2.29}$$

As for ES and ECA, we threshold the similarity matrix $\boldsymbol{S}^{\text{Pearson}}$ at some value to obtain an adjacency matrix for the construction of functional climate networks.

2.3. Spatially embedded networks

In spatially embedded networks such as functional climate networks, nodes are additionally characterized by a location in some physical space. Therefore, functional climate networks do not only have a specific network topology but can also be analyzed with geometric approaches. In the following, we introduce a pair of network measures which are solely defined for spatially embedded networks. Second, we discuss a comprehensive framework which corrects for systematic biases due to heterogeneous node placement in spatially embedded networks. Third, we discuss a scheme that accounts for boundary effects which we employ when we analyze a network in a confined region.

2.3.1. Network measures

In spatially embedded networks, we can measure distances between nodes [15, 79]. Accordingly, we can quantify the distance specific edges are covering. Next to possible global measures such as the global mean link distance or the maximal link distance

present in a network, its node-wise pendant, the *average link distance* d is a popular measure. For a node i, the average link distance is defined as [30]

$$d_i = \frac{1}{k_i} \sum_{j=1}^{N} a_{ij} d_{ij} \tag{2.30}$$

with d_{ij} representing the distance between the nodes i and j.

In addition to these distance-based measures, a few studies quantifying the geometry of a network have been conducted [79, 80]. Additionally, there have been introduced edge directionality measures which are one approach to quantify the directional tendency of edges adjacent to a node [80]. A detailed analysis of such geometric network measures and their application to study complex networks embedded in spherical geometry is presented in chapter 3.

2.3.2. Geometry correction

In spatially embedded networks where nodes represent heterogeneously sized entities, systematic biases have been observed [81–84]. Here, we shortly address two differently sourced biases and their corresponding correction schemes.

In some types of networks, the probability that an edge is present in the network depends on the physical distance between nodes. This is, in particular, the case for complex networks in Climatology and the corresponding functional climate networks which have been introduced in the previous section (section 2.2). Nodes in functional climate networks are commonly located on a regular spherical grid with constant latitudinal and longitudinal distance but heterogeneous spatial node density [82]. Specifically, in regular spherical grids, the density of nodes increases with the inverse of the cosine of the latitude [81]. In a scenario where nodes are more likely connected if their distance in the physical space is small, nodes around the poles will automatically have a higher number of connections to their direct neighbors than nodes located around the equator . To account for such biases, Heitzig et al. [82] have developed a comprehensive framework that is based on node splitting and twin merging. Note that this framework is not exclusively useful in the context of functional climate networks but also for complex networks in neuroscience, where similar biases have been observed [82].

To account for differently sized regions that are represented by a single node in a network, Heitzig et al. [82] have proposed to use tailored node weights which are based upon imaginary mergers or splittings of nodes to resample the node placement and compensate the original misplacement. This concept is commonly known as *node splitting invariance (n.s.i.)*. As an effect, the authors have suggested weighing each node in a climate network by the fraction of the corresponding area each node represents. Accordingly, they have proposed to use node weights which are proportional to the cosine of the latitude ϕ.

The corrected *n.s.i-degree*, therefore, reads

$$k_i^{nsi} = \sum_{j=1}^{N} w_j a_{ij}. \tag{2.31}$$

where $w_j = \cos \phi_j$ symbolizes the respective node weight.

Along with this definition of the n.s.i.-degree, the authors have introduced a comprehensive scheme to construct n.s.i.-corrected network measures and came up with definitions of the most popular network measures (such as the clustering coefficient and betweenness). We utilize n.s.i. network measures in all chapters where we study complex climate networks and refer to the respective network measure disregarding the term *n.s.i.* to increase readability. We only use the uncorrected versions if we explicitly mention doing so.

An additional issue in networks with distance-dependent linkage occurs at the boundaries of the study area. Nodes located close to a boundary of a study area have a reduced number of neighboring nodes in a small physical distance [85]. Therefore, we expect an elevated average link distance and a smaller degree for those nodes in such networks. According biases have been intensively studied by Rheinwalt et al. [85]. In this study, the authors have not only proposed an algorithm to account for the corresponding bias but have also subsequently released a software package for python which we have used for the work presented in the thesis (`https://github.com/Rheinwalt/spatial-effects-networks/blob/master/sern.py`).

Their algorithm is based on the construction of surrogate networks that explicitly conserve the link distance distribution of the whole network. In the surrogate networks, nodes i and j are connected based on the probability of the corresponding link distance in the link distance distribution of the original network. To maintain the overall link density, this probability is adapted considering the number of all possible links of this length based on the node locations. In an application, this procedure is repeated for a large number (at least 100 times) to construct an appropriate number of surrogate networks. To finally obtain the boundary-corrected network measure for each node, Rheinwalt et al. have proposed to subtract the mean of the respective network measure in the surrogate networks from the value in the original network.

To this end, we mention that the boundary correction algorithm can also be used to correct for heterogeneous node placement in climate networks as the differing node density is reflected in the node distance distribution which scales the linking probability in the surrogate networks. Due to the relatively large computational resources which are needed for constructing the surrogate networks, utilizing appropriate node weights is normally the preferred choice when boundary effects are negligible.

2.4. Construction of networks in Computational Social Science

Social platforms, such as Facebook, Twitter, or Instagram are often originally organized as networks. Through the affordances of these platforms, people can uni- or bidirectionally engage via following or friendship connections and, therefore, naturally form a network [35]. Next to these obvious network representations, there are multiple other ways to abstract data into a network representation. In chapter 7 and chapter 8 we pursue two different approaches, which we shortly outline in the following.

In social networks, friendships and follower relations can have heterogeneous meanings as people often have many different contacts with varying interaction frequency. To construct a representation of the effective interaction network, we, therefore, do not study the social network (e.g. the friendship or follower network) itself but use the dynamic interactions as a definition of an edge in a network. Using this strategy, nodes in the network are still the different users but the network represents the temporally changing and potentially weighted interaction strength between user pairs. Thus, an inactive user with many connections might be a node with only a few edges in the interaction network, whereas an active user with only a few connections in the static network can have all connections represented with large weights in the interaction network.

Next to these direct user-user interactions, in which we normally disregard the exchanged content, we can construct networks based on the shared content. In this case, we can identify URLs, hashtags, or even specific content such as a video as a node. Edges in this context for example refer to the co-occurrence (under a post or within some time interval) of two of such entities. The resulting content networks are somewhat disconnected from the actual user interaction but reflect the popularity and relation of different topics.

We refer to chapter 7 and chapter 8 where we conduct comprehensive data analyses of two data sets using the presented rationales.

Chapter 3.

Edge directionality measures

3.1. Introduction

To access the information hidden in network representations of complex systems, various network measures and methodological approaches have been introduced in chapter 2. In addition to the traditional topological analysis, we, here, present geometrical network measures to extract information from the spatial arrangement of edges. By utilizing the rationale from previous work [80] where the authors have proposed a geometric perspective on complex networks by introducing the concept of edge anisotropy, we further develop and generalize the toolbox of such measures to study networks embedded in spherical geometry.

To show the variety of insights that can be accessed using edge directionality measures, we first tie in with the concept of edge anisotropy which has been formulated by Molkenthin et al.[80]. As a second step, we slightly enlarge the number of the so-called edge directionality measures and present a framework to compute such measures in real-world networks which are commonly embedded in spherical geometry. We further show, that geometric network measures are prone to systematic biases and demonstrate a scheme to correct for corresponding effects. Finally, we illustrate the potential of geometric network measures by studying three different climate networks, an air transportation network and the global trade network from 2009.

The results presented and the figures shown in this chapter are based on publication P_1 (*Wolf, F., Kirsch, C. and Donner, R. V. (2019): Edge directionality properties in complex spherical networks. Physical Review E 99.1, 012301*). We thank APS for the kind permission to reuse/adapt the content and figures.

3.2. Geometric network measures

In principle, geometric network measures quantify spatial arrangement of edges that are adjacent to a distinct node. This consideration adds another layer of information to the sole topological network analysis. In spatially embedded networks, nodes are not only characterized by their topological attributes but also by a physical location. Therefore, we can attribute a unit vector $\vec{e}_{ij}$ pointing from node i towards another node j along an edge if both nodes are connected in the studied network.

Following this rationale, the authors of a previos study [80] have suggested quantifying the edge anisotropy as

$$r_i = \frac{1}{k_i} \left| \left| \sum_{j=1}^{N} a_{ij} \vec{e}_{ij} \right| \right|. \tag{3.1}$$

To not only analyze a scalar value indicating anisotropic edge placement, we suggest to further discuss the *mean edge direction* [80],

$$\vec{r}_i = \frac{1}{k_i} \sum_{j=1}^{N} a_{ij} \vec{e}_{ij}, \tag{3.2}$$

which illustrates the geometrical imbalance of the edge directions. Note that $r_i = ||\vec{r}_i||$.

The definition of the weighted mean edge direction and the weighted edge anisotropy follows directly by replacing the adjacency matrix elements a_{ij} with the weight matrix elements w_{ij}.

To apply these edge directionality measure when analyzing directed networks, Molkenthin et al. [80] introduced the *in- and out-edge anisotropy* as

$$r_i^{in} = \frac{1}{k_i^{in}} \left| \left| \sum_{j=1}^{N} a_{ij} \vec{e}_{ij} \right| \right|, \tag{3.3}$$

$$r_i^{out} = \frac{1}{k_i^{out}} \left| \left| \sum_{j=1}^{N} a_{ji} \vec{e}_{ji} \right| \right|. \tag{3.4}$$

The *in- and out-mean edge direction* are computed accordingly.

3.3. Anisotropy in spherical geometry with homogeneous areas of representation

Nodes in climate networks are characterized by their topological properties as well as a latitude ϕ_m and longitude λ_m. The latter specify the location on a sphere that is embedded in three-dimensional space. An edge pointing from node i to node j can, therefore, be assigned the course angle α_{ij} (we present a detailed derivation of the analytic expression for course angles in Appendix A)

$$\alpha_{ij} = \arccos c_{ij} \tag{3.5}$$

with

$$c_{ij} = \frac{\cos\phi_i \sin\phi_j - \cos(\lambda_i - \lambda_j)\cos\phi_j \sin\phi_i}{\sqrt{1 - (\cos(\lambda_i - \lambda_j)\cos\phi_i \cos\phi_j + \sin\phi_i \sin\phi_j)^2}}. \tag{3.6}$$

The course angle α_{ij} which has originally been introduced to quantify the angle between the geodetic north and the course of a moving ship is, by definition bounded between $\alpha_{ij} \in [0, \pi]$ and can be attributed to an edge connecting node i and node j. To distinguish directions with eastward and westward component we further define $\beta_{ij} = \alpha_{ij}$ for the former and $\beta_{ij} = 2\pi - \alpha_{ij}$ for the latter.

Finally, for the calculation of the mean edge direction, we locally neglect the curvature of the Earth's surface and set

$$\vec{e}_{ij} = \begin{pmatrix} \sin\beta_{ij} \\ \cos\beta_{ij} \end{pmatrix} \tag{3.7}$$

where the second coordinate denotes the northward component in this definition. Furthermore, we name the angle δ_i between the mean edge direction $\vec{r}_i$ (Eq. 3.2) and the geodetic north the *local mean angle*.

3.4. Anisotropy in spherical geometry with heterogeneous areas of representation

In networks where nodes represent differently sized regions or entities, network measures can be biased. To account for such inaccuracies the framework of *node-splitting invariance* (n.s.i., see chapter 2) has been introduced by the authors of [82] and further specified in subsequent studies [83, 84].

While the corresponding biased degree and local clustering coefficient values in climate networks have been discussed by the authors of [82], we address an analogous bias in edge anisotropy and local mean angle. In accordance with the already introduced framework, we define the corrected local anisotropy by

$$r_i^{AOR} = \frac{1}{\sum_j w_{ij}^{AOR}} \left\| \sum_{j=1}^{N} w_{ij}^{AOR} \vec{e}_{ij} \right\|, \tag{3.8}$$

which can, again, be generalized for weighted networks where we suggest a multiplicative combination of intrinsic and *AOR*-weights (AOR is an abbreviation for *area of representation)*. Here, we do not utilize the term *n.s.i.*-weights as we employ edge properties. We, therefore, have to introduce specific edge weights in contrast to the approach of Heitzig et al. [82] who have suggested the use of node weights.

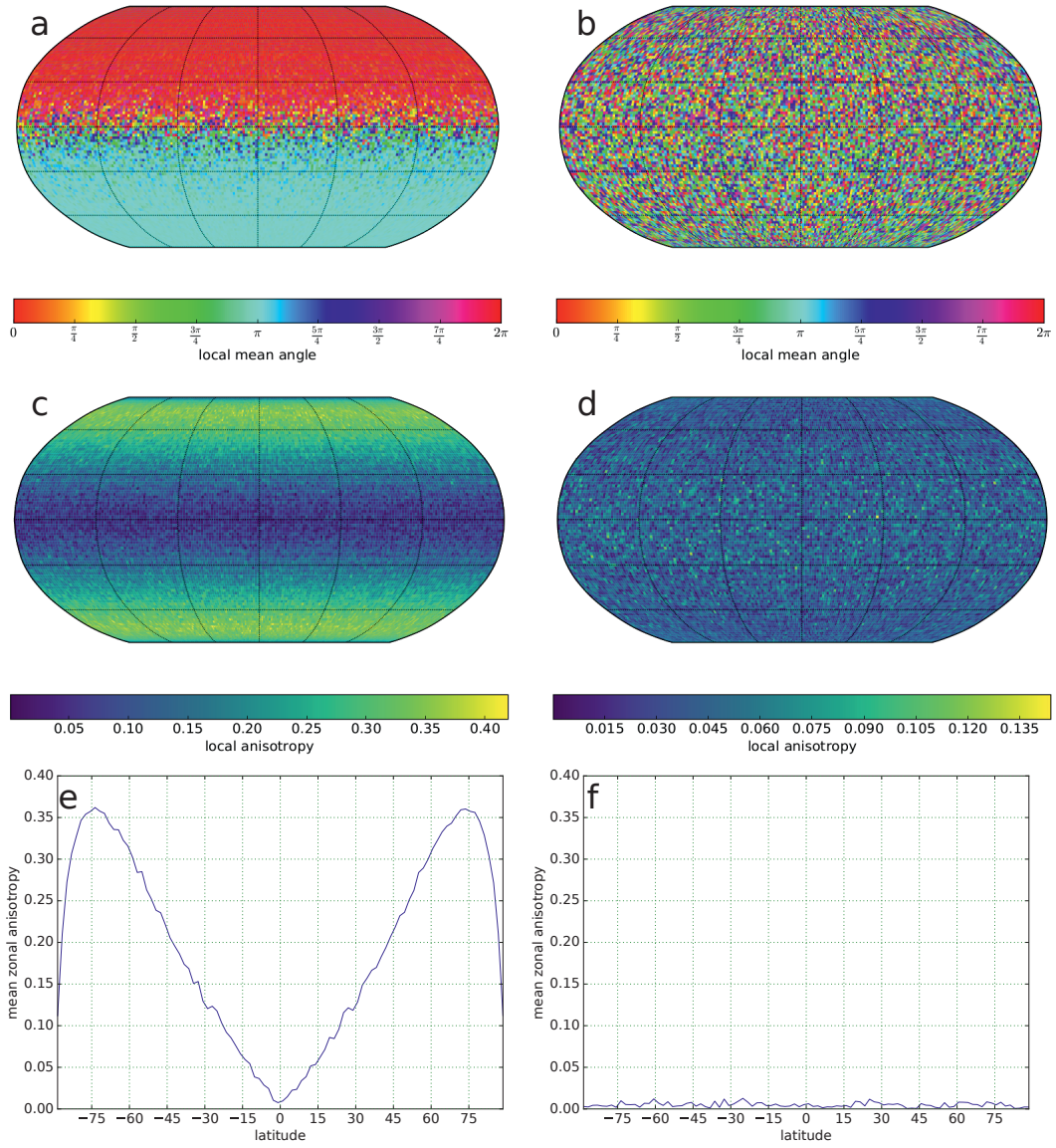

Figure 3.1.: Local mean angle (a,b), local anisotropy (c,d) and zonal mean local anisotropy (e,f) of the benchmark network. The left column (a,c,e) shows the uncorrected values and the right column (b,d,f) the corresponding corrected values.

3.5. Bias correction for edge directionality measures

To investigate possible systematic biases in edge directionality measures, we study a synthetic network that is embedded in a regular spherical grid where heterogeneous node placement naturally arises. Here, we utilize a rotationally and translationally symmetric benchmark network which we, in line with previous work [82], construct utilizing the linking probability

$$p(d_{ij}^{\alpha}) = \min(1, \exp(0.4 - 0.09 d_{ij}^{\alpha})) \tag{3.9}$$

depending on the angular distance d_{ij}^{α}. This imitates the preference towards smaller link distances which is observed for many correlation based climate networks. The parameter setting $(0.4, 0.09)$ assures a link density of $\rho = 0.035$ which is of the order commonly used in climate networks. Our results are based on a network with $N = 18432$ nodes on a spherical grid with a resolution of $1.875° \times 1.875°$ latitude and longlitude.

Although we follow the described random linkage mechanism, we find systematic gradients in local mean angle (Fig.3.1a) and local anisotropy (Fig.3.1c,e).

Figure 3.1a features the hemispheric splitting of the local mean angle to values around 0 (or 2π) on the Northern Hemisphere and values around π for the Southern Hemisphere with a transition region in between. The local anisotropy, in turn, increases from zero at the equator to a maximum around $\pm 75°$ latitude followed by a marked decrease around the poles on both hemispheres.

Such a systematic derivation from random fluctuations is no feature of the benchmark network itself but can be attributed to the spatial embedding of the network on the regular spherical grid.

On a regular spherical grid, where the angular distance between nodes is kept constant, the node density increases with the inverse of the cosine of the latitude which can be proven straightforwardly by considering the area of a spherical rectangle $A_{\square}$ associated with a node at latitude ϕ_i. The size of the spherical rectangle is proportional to the difference between two spherical caps cut at $\phi_i - \frac{1}{2}\Delta\phi$ and $\phi_i + \frac{1}{2}\Delta\phi$ where $\Delta\phi$ denotes the difference between the neighboring latitudes ϕ_i and ϕ_j.

Following this notation, we compute

$$\begin{aligned} A_{\square} \propto A_1 - A_2 &\propto \sin\left(\phi_i + \frac{1}{2}\Delta\phi\right) - \sin\left(\phi_i - \frac{1}{2}\Delta\phi\right) \\ &= 2\sin\frac{\Delta\phi}{2}\cos\phi_i \propto \cos\phi_i \end{aligned} \tag{3.10}$$

and conclude that the area of spherical rectangles decreases with the cosine of the latitude.

This implies that the distance between nodes depends on the direction in which we measure the distance. Let us consider a node on the Northern Hemisphere: in the

vicinity of the node, we find a higher link density and smaller distances between nodes in the north than towards the south. Therefore, distance-dependent linkage schemes automatically induce a preference towards the poles resulting in a systematic bias in the mean angle and an increase in the local anisotropy. The marked decrease of the local anisotropy at the poles can also be explained by the distance-based linking scheme. Due to the decrease of $p(d_{ij}^{\alpha})$ to $\frac{1}{e}$ at $d_{ij}^{\alpha} \sim 15°$ latitude, a critical portion of edges cross the polar region and are connected to nodes of similar latitude and, hence, cause the decrease for values $\gtrsim |\pm 75°|$.

To correct for this bias, we suggest utilizing specific edge weights to compensate for the systematic deviation. In climate networks, the choice $w_{ij}^{AOR} = \cos\phi_i$ has been shown to correct for the biases of topological network measures such as degree or clustering coefficient [82]. Correspondingly, we suggest to weigh edges pointing from node i to node j with the AOR-weight $w_{ij}^{AOR} = \cos\phi_i$ to obtain the unbiased versions of the edge directionality measures which we show in Fig. 3.1b,d,f. Confirming the success of the bias correction, Fig. 3.1b shows uniformly distributed mean angles in $[0, 2\pi]$ and Fig. 3.1d,f demonstrate that the local anisotropy exhibits solely random fluctuations with an amplitude of 10^2 lower than in the biased version.

Having set up the theoretical framework to analyze networks using the concept of edge directionality, we continue by studying several real-world networks in the following. Thereby, we illustrate the broad applicability of the introduced edge directionality properties.

3.6. Edge directionality in climate networks

3.6.1. Data and network construction

In this section, we demonstrate the capability of edge directionality measures to study climate networks. Initially, we study a surface air temperature (SAT) data set from the ECHAM 6.1 AquaControl simulation which was performed within the TRACMIP coordinated experiment [86]. Here, the authors have presented comprehensive simulations, where an idealized planet without landmasses is studied. This allows for a detailed analysis of atmospheric processes without possible disturbances due to landmasses. The studied aquaplanet features a thermodynamic slab ocean with interactive sea-surface temperatures and air-sea coupling as well as follows present-day insolation [86, 87]. An implemented hemispherically asymmetric, but zonally symmetric meridional heat transport shifts the position of the inter-tropical convergence zone (ITCZ) northwards and thereby causes asymmetric circulation cells [86]. The simulation covers 30 years with monthly values and a spatial resolution of approximately $1.875° \times 1.875°$ latitude and longitude. Note that we present a comprehensive study of multiple different model simulations performed within the TRACMIP experiment in chapter 6.

Besides, we study global monthly precipitation sums and SAT averages between 1979 and 2016 from the ERA-Interim reanalysis data set, which are available at a spatial resolution of $2.5° \times 2.5°$ [88].

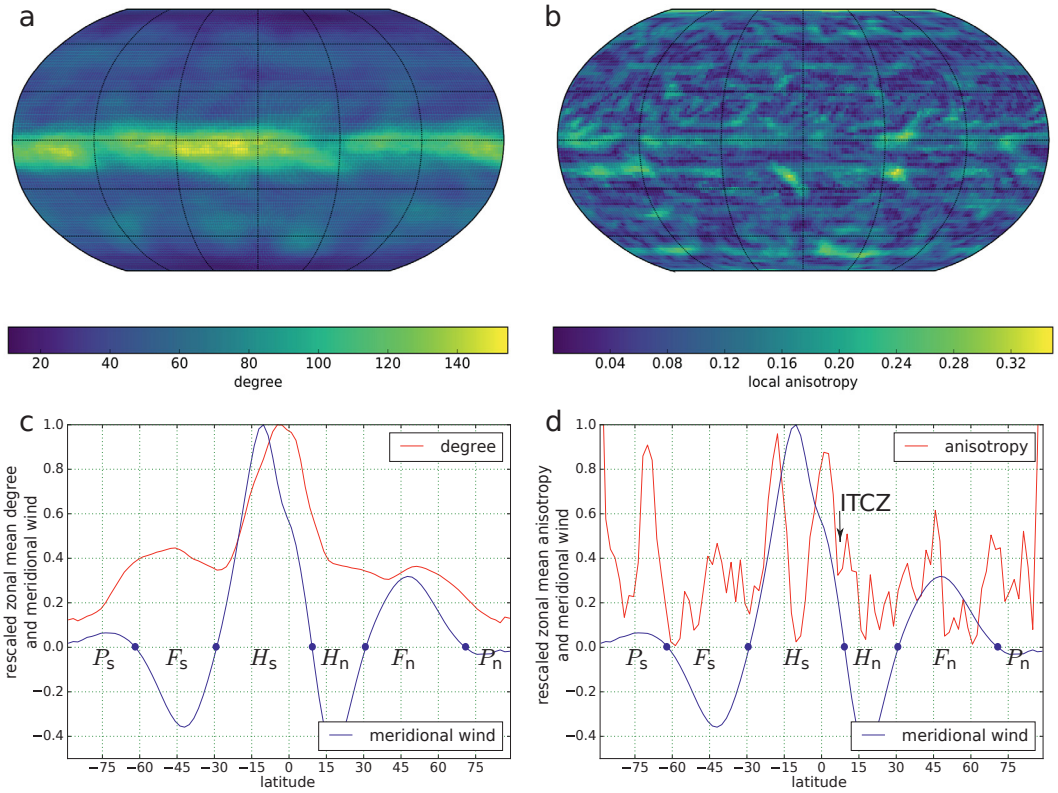

Figure 3.2.: Comparison between degree and local anisotropy in a climate network. (a) shows the n.s.i. degree, (b) the local anisotropy. The meridional wind with marked circulation cells is shown together with the rescaled zonal mean degree (c) and the rescaled zonal mean local anisotropy (d). P,F,H indicate the circulation cells, subscripts the hemisphere.

In both cases, we construct climate networks as described in chapter 2 by using the respective anomaly time series and Pearson correlation as the similarity measure. We finally obtain the adjacency matrix by thresholding the similarity matrix at some value to achieve a link density of $\rho = 0.005$ which is a typical value in global climate networks with similar spatial resolution [28, 30].

3.6.2. SAT aquaplanet climate network

Figure 3.2a features the degree of the SAT aquaplanet climate network which is zonally symmetric due to the absence of landmasses and corresponding symmetric dynamics. Whereas there is not much complex structure observable in the global degree pattern (Fig. 3.2a), Fig. 3.2c reveals a close relation between the zonal mean degree and the meridional wind of the hemispherically asymmetric circulation cells. The cells themselves are not directly observable in the degree field but we observe some relation between maxima of the meridional wind and the degree which we

suspect to arise from the transportation of temperature anomalies by low-level winds. This has been described also in a previous study [80].

In contrast to the degree, the local anisotropy which is shown in Fig. 3.2b exhibits complex structures and allows for a more detailed analysis of the climatic features.

First, we observe that the northward-shifted ITCZ (4.6°N [86]) is characterized by a low anisotropy canal in Fig. 3.2b and a local minimum in the zonal mean local anisotropy. As implied by the name *convergence zone*, winds merge from northern and southern direction resulting in links into both directions and, thus, a relatively low local anisotropy. This observation is supported by the other local minima of the local anisotropy for the edges of the circulation cells on the Southern Hemisphere. On the Northern Hemisphere, we observe trade wind-related structures in Fig. 3.2b and generally smaller circulation cells with weaker meridional wind maxima. This leads to weaker signatures of circulation cell edges in the local anisotropy as the Southwest-to-Northeast-orientated trade wind imprints blur the edges of the Ferrel (F_n) and Hadley cell (H_n).

Second, we find a specific double band (or double peak in the zonal mean) structure of elevated anisotropy related to the southern Hadley cell H_S which ranges from ~ 4°N to ~ 30°S (Fig. 3.2a,c). Implied by the coinciding high node degree covering the southern Hadley cell, we suspect that this might be a characteristic feature of the edge anisotropy in densely connected parts of a network. Due to the densely connected nodes, we find a minimum of the anisotropy in the center of the Hadley cell and edge directions pointing towards the interior when we analyze the nodes more towards the edges of the circulation cells. At the boundary of the circulation cells, finally, the edge direction completely changes which leads to the already described edge anisotropy minima between the cells.

Third, the most striking similarity is, in line with the observation from the southern Hadley cell, a relative minimum of the local anisotropy in the center of the circulation cells and maxima slightly shifted to the respective mean boundaries which can be observed for almost all circulation cells (except the southern Ferrel cell). A previous study [80] already mentioned that large degree values and low local anisotropy are related to fast directed flow in model systems, which we also expect in the center of the circulation cells. In turn, at the boundaries either moderately diverging or moderately converging winds transport anomalies in a directed manner towards or from the circulation cell center which might also explain large anisotropy values close to the boundaries. However, as we here exclude interannual shifts of the whole system of the circulation cells and the ITCZ and disregard perturbations due to the Coriolis force or atmospheric waves, deviations are to be expected and are indicated by the large general variability of the anisotropy along the latitudes.

3.6.3. SAT and precipitation real-world climate networks

For the SAT and the precipitation network, we show the local anisotropy and the mean edge direction for nodes with $r_i > 0.25$ and $k_i > 10$ in Fig. 3.3.

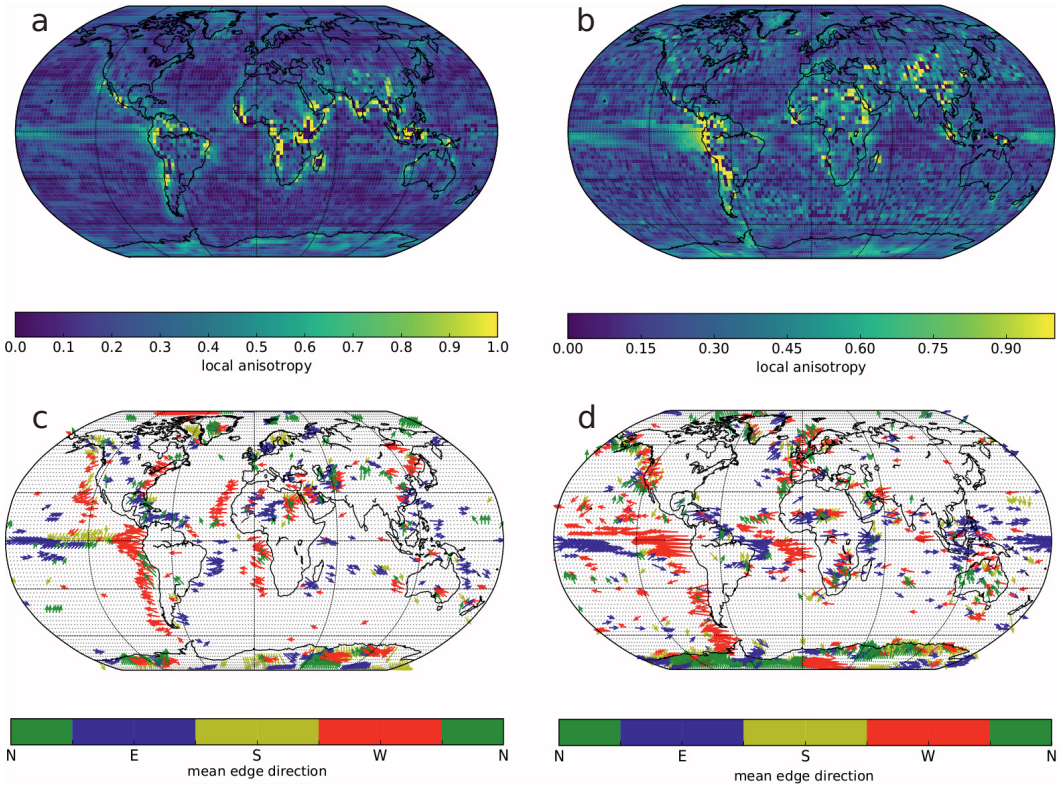

Figure 3.3.: Local anisotropy (a,b) and mean edge direction (c,d) of the real-world SAT (a,c) and precipitation (b,d). The mean edge direction is color-coded by the cardinal direction and only shown for nodes with $r_i > 0.25$ and $k_i > 10$.

For the former (Fig. 3.3a,c), we mainly find elevated values of edge anisotropy along the coasts, which is caused by the abrupt change in heat capacity at the coasts leading to different temperature dynamics over land and ocean. At the boundary of the regions with different dynamics, we, therefore, find a flip in the mean edge direction. This gets amplified by deflected (trade) winds which transport temperature anomalies. In addition, we predominately show the mean edge direction over the ocean in Fig. 3.3c as we observe larger values of the degree over the ocean due to elevated spatial temperature persistence. The El Niño–Southern Oscillation (ENSO) which is one of the largest temperature variabilities of the Earth's climate system is the second feature highlighted in the SAT network. The elevated degree values and highly connected nodes in the South Pacific originate from the well-known sea surface temperature anomalies occuring with varying frequency (2 to 7 years) in the Pacific.

In the precipitation network (Fig. 3.3b,d), we identify two main characteristics. On one hand, we again observe a marked region of elevated local anisotropy values related to the ENSO. Interestingly, we find a double band structure similar to the signature of the zonal mean anisotropy of the southern Hadley cell in the aquaplanet network confirming that double bands in the edge anisotropy are characteristic features of

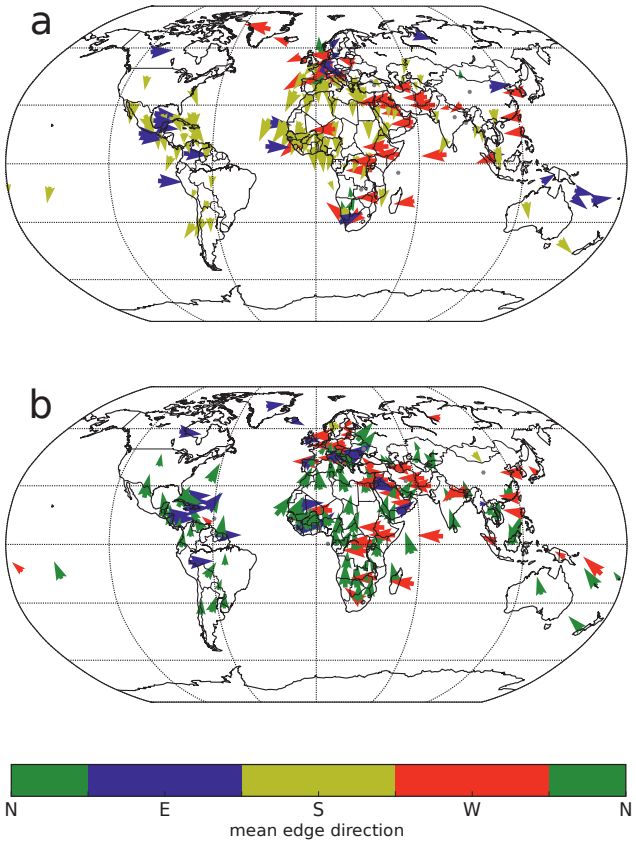

Figure 3.4.: Mean edge direction of the import (a) and export (b) trade network. The mean edge direction is color-coded by the cardinal direction and only shown for nodes with $r_i = 0.3$ and $k_i = 25$.

large regions of high interconnectivity. A similar structure with lower amplitude and a smaller spatial extent can be identified that coincides with the position of the Atlantic Niño between South America and Africa [89]. The identification of both is remarkable as we, in this part, solely utilized the geometry of the network edges and excluded traditional topological measures.

3.7. Edge directionality in trade networks

To complement the previous study of the climate network, we investigate the global trade network with nodes representing different countries. Such networks have already been analyzed in previous studies utilizing topological approaches [13, 90].

The studied network is based on the year 2009 world trade network, which we construct from the Eora multiregional input-output database [91]. The $E = 7,043$ edges in the network represent the directed and weighted trade of goods between $N = 186$ countries, each symbolized by a node at its geographical center.

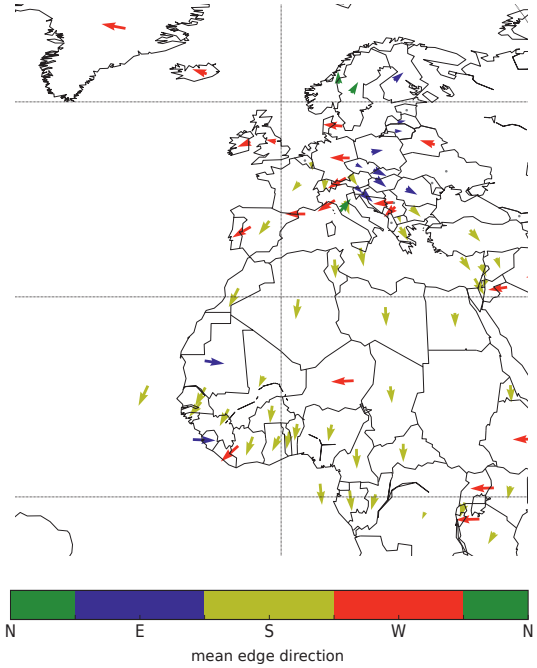

Figure 3.5.: Snippet from Fig. 3.4 showing the mean edge direction of the import trade.

The directed global trade network is characterized by relatively large anisotropy values and dense interconnectivity (link density of $\approx 20\%$). To illustrate these features, we show the incoming edge direction (colored according to their cardinal direction) referring to the import in Fig. 3.4a and the corresponding outgoing and export-related edge direction in Fig. 3.4b for all nodes which exceed $r_i = 0.3$ and $k_i = 25$. This criterion is met by 171 of the 186 countries illustrating the high entanglement of the world's trade.

By comparing the import and the export directionality of the network, we can distinguish between different actors of the global trade. On a global scale, the import and export mean edge direction differ and are easily discriminated against. In the detailed analysis of the single nodes representing different countries, we mainly find two classes: either export and import align or exhibit opposite edge direction. We suspect this difference to arise from the position in global supply chains. Whereas some countries are at intermediate positions (e.g. Canada, receiving import from Asian countries, exporting to the east) other countries stack many intermediate goods to the final product and, therefore, exhibit opposite export and import direction (e.g. Australia).

Another local anisotropy feature in the trade network is caused by the signature of the European export. As an influential actor in the world's economy, the European Union accounts for approximately one-third of the export trade worldwide. Especially North and West African, Middle Eastern, and South American countries receive many goods from the EU resulting in import edge directions pointing away from the

European Union (see Fig. 3.5). Also in Fig. 3.5, we observe mean edge directions flipping at the border between countries which joined the EU pre- and post the EU enlargement in 2004. Whereas this could be a relict from older trade agreements, this particular edge directionality pattern (western European countries pointing westward, eastern European countries vice versa) can also be related to the large trade volume in a confined region which automatically leads to diverging import edge directions.

3.8. Edge directionality in air transportation networks

Finally, we analyze an air transportation network to illustrate the potentials of geometric network measures to analyze such complex systems. Air transportation networks have also previously been analyzed using network approaches [92, 93].

Here, we study the airport and route datasets from the *OpenFlights* database (`https://openflights.org/data.html`) containing $N = 7,184$ airports (interpreted as nodes) connected via $E = 67,663$ weighted, directed edges, each representing different flight routes.

In line with the previous analysis, we again only show airports with an elevated degree and anisotropy ($r_i > 0.3$ and $k_i > 30$) in Fig. 3.6 and, hence, neglect smaller airports (in terms of flight volume).

In Figure 3.6a, mean edge directions around the Pacific pointing away from the ozean indicate that only a comparably low number of flights cross the Pacific Ocean. Whereas this water body is avoided due to its size, we cannot generalize this observation to the other oceans, as most North and South American East coast airports exhibit mean edge direction pointing towards the East. This also implies that larger coastal airports do not mainly serve domestic flights because the resulting mean edge direction had to point towards the inner part of the continent in this case.

Another feature of the analyzed network is the marked southern-pointing mean edge direction in most parts of Europe. We suspect that this is a result of the placement of Europe on the Northern Hemisphere in combination with the high connectivity to all parts of the world, which sum up to a purely southern pointing mean edge direction in the outgoing flights.

Finally, we can also identify subsystems of locally organized structures utilizing geometric network measures. In the present network, we, accordingly, find the Asian air traffic with a distinct edge directionality pattern as a relevant subsystem. Due to the large inner-Asian flight volume, we find converging edge directions pointing towards the center of China at airports in many countries surrounding the Chinese mainland (see Fig. 3.6b).

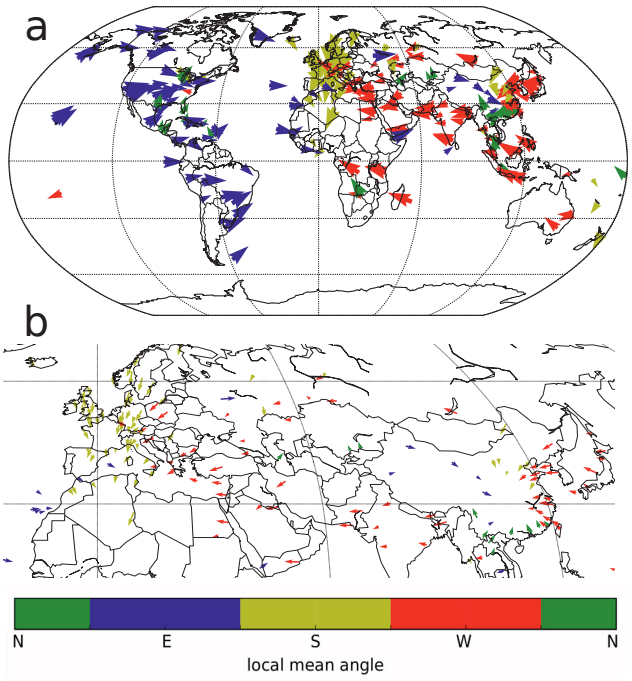

Figure 3.6.: Global (a) and Eurasian (b) mean edge direction of the air transportation network. The mean edge direction is color-coded by the cardinal direction and only shown for nodes with $r_i = 0.3$ and $k_i = 30$.

3.9. Summary

Geometric network measures can complement traditional topological network analysis. To confirm this assumption, we have initially presented a comprehensive set of geometric network measures. Motivated by the study of real-world networks which are commonly embedded in spherical geometry, we have generalized the edge direction properties accordingly. To avoid geometry-induced biases in the network measures, we have discussed a systematic bias that is induced by heterogeneous node placement. To correct for this bias, we have employed a corresponding correction scheme and, thereby, set up a framework to apply edge directionality measures to investigate real-world networks.

In climate networks, we have shown that geometric network measures reveal large-scale phenomena like the El Niño–Southern Oscillation or the Atlantic Niño and exhibit complex patterns indicating trade wind directions and circulation cell centers and edges.

In trade networks, edge directionality measures have confirmed and visualized the importance of the EU export trade and have characterized countries as endpoints or in-between members of international supply chains.

Finally, the analysis of an air transportation network has displayed different air transportation subsystems and has indicated the influence of big water masses on the global organization of air transport.

Part II.

Complex networks in Climatology

In the second part of the thesis, different properties of the Earth's climate system using complex networks are examined. First, the caveats of network construction based on event synchrony in chapter 4 are addressed. Using two case studies focusing on distinct features of the South American Monsoon System, the characteristics of Event Synchronization and Event Coincidence Analysis are highlighted. With the help of Event Coincidence analysis, a study on the temporal evolution of heavy rainfall event patterns related to East Asian Summer Monsoon is conducted in chapter 5. In chapter 6, the presented work contributes to the understanding of global rainfall distribution associated with the intertropical convergence zone by analyzing sea surface temperature anomalies of an ensemble of idealized aquaplanet models. For this work, a distinct approach for constructing the according functional climate networks is utilized.

Chapter 4.

Event synchrony measures for functional climate network analysis

4.1. Introduction

Studying extreme events, their co-occurrence and their changing characteristics in the context of ongoing climate change is an integral part of climate impact research. This is not only motivated by the impact of extreme weather on the economy of regions or countries but also by their threat to people's lives [21, 22, 94]. To support the effort of investigating extreme events, multiple studies have employed Event Synchronization (ES) [33, 95] as a method to estimate the similarity between event time series. The results of this analysis can be used to construct functional representations of statistical interrelations between (cascades of) extreme events [5, 34, 70]. Utilizing a dynamic local coincidence interval, the ES can capture event synchrony at different time scales [5, 24] but inherits a systematic problem to capture clustered events [5, 77, 78]. While there has been suggested a correction scheme to correct for the resulting bias in functional climate networks [5, 24, 96], other authors have introduced Event Coincidence Analysis (ECA) [73] as a powerful alternative that is not susceptible to the mentioned clustering bias.

In this chapter, we study the already introduced event synchrony measures ES and ECA (see chapter 2) regarding their capabilities to construct and study climate networks. To comprehensively compare and contrast the ES and the ECA, we apply both methods in combination with and without the correction scheme to study the South American Monsoon System (SAMS) [29, 55, 96–98].

Accordingly, we first shortly introduce the main drivers and components of the SAMS, specify the parameter settings, and comment on the respective data preprocessing before we conduct two case studies highlighting the features and possible shortcomings of both methods. In particular, we investigate differently caused biases and a corresponding correction scheme in the first case study whereas we highlight the systematically differing potentials of the different event synchrony measures to capture cascading events across scales in the second.

The results presented and the figures shown in this chapter are based on publication P_2 (*Wolf,F., Bauer, J., Boers, N., Donner, R.V. (2020): Event synchrony measures for functional climate network analysis: A case study on South American rainfall*

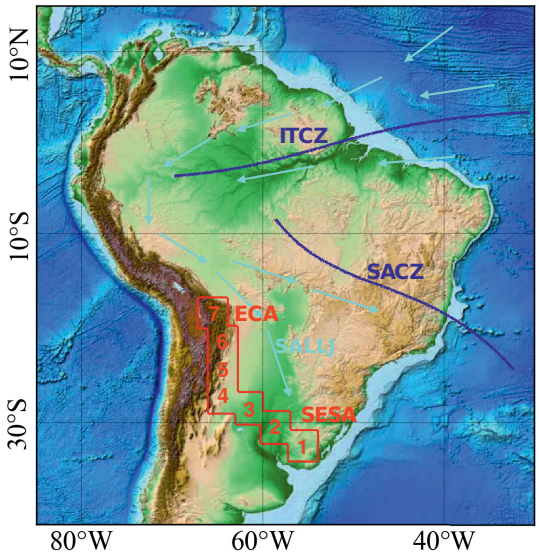

Figure 4.1.: Topography of South America and key features of the South American Monsoon System (SAMS), including typical wind directions (light blue dashed arrows) and the South American Low-Level Jet (SALLJ). The climatological positions of the Intertropical Convergence Zone (ITCZ) and the South Atlantic Convergence Zone (SACZ) are shown by dark blue lines. The red boxes illustrate the parcellation of the study area into 7 boxes to track the propagation of extreme precipitation events (see text).

dynamics. Chaos: 30.3., 033102). We thank AIP Publishing for the kind permission to reuse/adapt the content and figures.

Note that we present an extension of the first case study (section 4.5) in Appendix B.

4.2. The South American Monsoon System

In this exemplary application of event synchrony measures, we study heavy precipitation events related to the SAMS [29, 55, 96–98]. Covering the time period between December and February, the SAMS is mainly fed by trade-wind driven moisture influx from the Atlantic ocean which converges at the ITCZ (see Fig. 4.1) [99]. The converging winds carry moisture, which is recycled over the Amazon basin towards the northern part of the Andes mountain range, where the winds get blocked and channeled southwards [84, 100]. From there, low-level winds distribute the moisture either towards South East Brazil (SEBRA) or (via the South American Low-Level Jet - SALLJ) South East South America (SESA) [101–105]. This distribution of moisture is highly influenced by the Rossby wave train phase [106] and results in the South American Rainfall Dipole between SEBRA and SESA as the most prominent rainfall variability in South America [97, 107].

4.3. Data and network construction

As mentioned in the introduction to this chapter, we mainly aim for a comprehensive comparison of ECA and ES especially in the context of potential biases of both methods resulting from the analysis of temporally clustered events in time series.

For our study, we utilize the rainfall data set from the Tropical Rainfall Monsoon Mission (TRMM, version 3B42 V7) [25], which is available on a spatial resolution of $0.25° \times 0.25°$ latitude and longitude covering the time between 1998 and 2015 with a temporal resolution of 3h. Specifically, we analyze the rainfall during the South American Summer Monsoon in the period from December to February (DJF) and conduct two case studies, one based on daily rainfall sums and one based on the raw 3-hourly data.

Subsequently, we consider rainfall to be *extreme* if the daily sum or the 3-hourly raw data exceeds the 90th (daily data) or the 98th percentile (3-hourly data). Given enough rainfall days, this results in a constant number of events in all time series (172 for the daily, 232 for the 3-hourly data). With only 4379 of the daily time series and of 5087 of the 3-hourly time series (out of in total 48400 time series) deceeding 90% of the respective maximum number of rainfall events (predominantly the continental shadow west of the Central Andes), we proceed by not treating possible dry spots differently.

Utilizing the stated definition of extreme events and specifying the ES and the ECA as the chosen similarity measures, we follow the described procedure (see chapter 2 for details), to construct functional climate networks. The parameters are chosen differently for the two case studies and are specified in the respective sections. In all figures showing the spatial pattern of different network measures we show the geometry corrected version of the respective network measure which we obtain using the algorithm from Rheinwalt et al. [85] (see chapter 2 for further details).

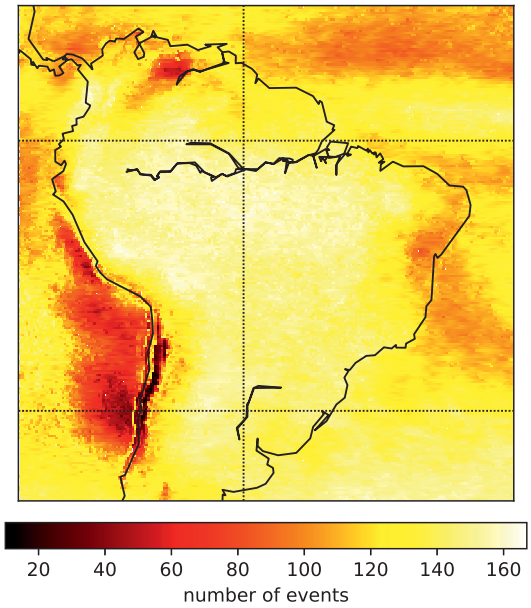

Figure 4.2.: Number of remaining events after declustering the obtained event sequences with initially a constant number of events for daily precipitation data (see text for details).

4.4. Declustering scheme for extreme event time series

In event time series, events are potentially temporally clustered and thereby induce a bias when applying ES in its original definition. This bias mainly arises from the collapse of the local dynamic coincidence interval (Eq. 2.17) to $\frac{1}{2}$ for subsequent events and, therefore, only allows for simultaneous coincidence (we consider a sequence of at least two subsequent events as an event cluster). This resulting bias and possible implications have been discussed in recent publications [77, 78]. One way to cope with this issue has been proposed and applied in several studies from Niklas Boers et al. (e.g. [5, 24]) and is based on the following correction scheme: before the application of ES we need to individually analyze each time series alone and only keep the first event of all event clusters in the time series.

Although the scheme corrects for the described issue of ES, we can further motivate its application as a tool for interpreting a sequence of subsequent events as a long persistent event [5]. Especially for analyzing precipitation time series where heavy rainfall events are often related to major weather systems, we can argue that there is also a physical meaning behind the correction scheme. In this context, we conclude, that it is also worth studying the effects of the correction scheme on networks constructed using the ECA, where subsequent events do not imply a systematic bias due to the definition of the global, static coincidence interval.

Depending on the overall strategy of constructing a functional climate network based on event synchrony, the application of the correction scheme can have a considerable effect on the corresponding similarity on which we define edges. Most

notably, we reduce the number of events in the time series by only keeping the first of all event clusters which directly influences the similarity values of ES and ECA.

Here, there are different ways to tackle this issue. In the studies where the correction scheme has been applied [5, 24] edges are defined by tailored significance testing utilizing surrogate time series with the respective combination of event numbers. Although this strategy is coherent and specified to the present issue, it does not conserve the serial dependency of extreme events. Another possible way to approach this problem is to include more extreme events step by step to even out event numbers, which appears to be computationally costly and incoherently sets event thresholds.

However, in the present study, we do not aim for revealing new climatic features of the SAMS. Therefore, we follow the most naive approach and neglect the existence of differing event numbers in the time series which we justify by relatively low changes in event numbers in most regions. In Fig. 4.2, we show the numbers of extreme events for time series with daily resolution in the respective grid cells which are initially influenced by possibly too little wet days (as described in the previous section) and then secondly by the correction scheme. In particular, Fig. 4.2 indicates decreased numbers of events in the Pacific ocean related to the continental shadow west of the Central Andes, the Atacama desert, the Orinoco Basin, and some regions in the northern Atlantic Ocean. In the analysis, we further investigate related network changes and especially discuss prominent pattern in the northern Atlantic ocean.

Summarizing, we utilize the described data sets along with the proposed event definition and calculate the similarity matrix $\mathbf{S}$ with and without the declustering scheme. We finally obtain the adjacency matrix by thresholding $\mathbf{S}$ to acquire a link density which we specify in the sections of the respective case study. For details on the network construction, see chapter 2. We also want to stress, that utilizing the uncorrected version of the ES in combination with employing a threshold for defining extreme events (as we do it in this chapter) is a reasonable strategy and frequently used for constructing climate networks [75, 76].

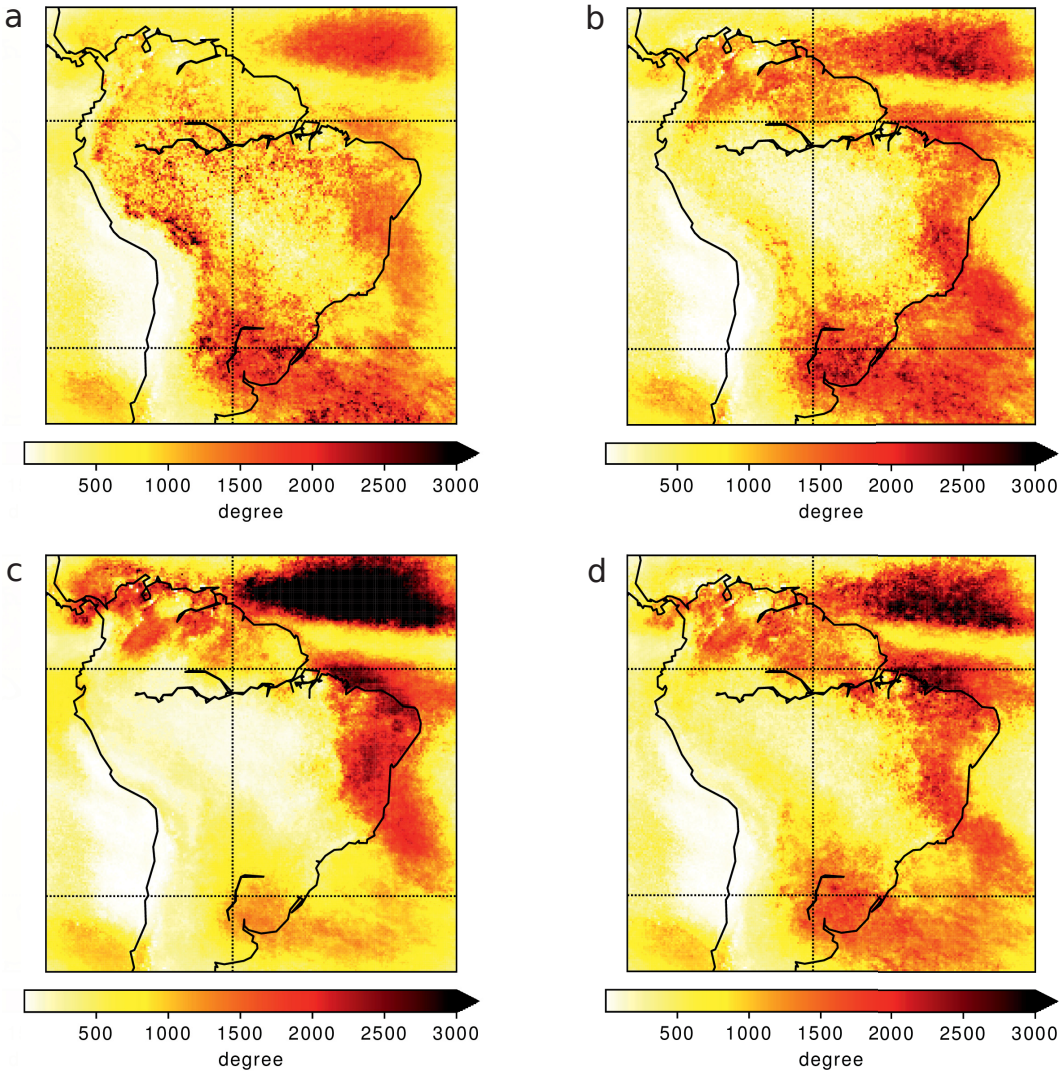

Figure 4.3.: Degree of the functional climate network representations of heavy rainfall events based on (a,b) ES and (c,d) ECA without (a,c) and with (b,d) the utilization of the correction scheme for temporally clustered events. For the two ES based networks, we set $\tau_{lm}^{ij} \leq 3$ days, while for ECA, $\Delta T = 3$ days. All networks exhibit a link density of 0.02.

4.5. Case study 1: Network pattern of the SAMS

4.5.1. Method intercomparison

Initially, we study the SAMS by constructing networks from the daily precipitation estimates. To tie in with previous studies, we set $\tau_{lm}^{ij} \leq \tau_{max} = 3$ days to avoid overly large (and unrealistic) coincidence intervals (Eq. 2.17 and Eq. 2.22) and thereby copy a previously employed setup of the ES [29] where the authors studied the SAMS without the application of the correction scheme. Subsequently, we threshold the similarity matrix $\mathbf{S}$ (Eq. 2.21 and Eq. 2.26) at some value to obtain a link density (Eq.2.1) of $\rho = 0.02$ which is of the common order for climate networks with comparable resolution [30, 57].

According to the already mentioned publication [29], we first apply the ES (for definition see section 2.2.1) without the utilization of the correction scheme. In the node degree (Eq. 2.31) of the resulting network (Fig. 4.3a), we identify the ITCZ as a low degree channel in the northern Atlantic. In addition, we recognize elevated node degree along the East Central Andes where the SALLJ transports moisture towards the south. In comparison to the moisture exit zone in SESA which is marked by high node degree values, SEBRA, as the second part of the South American Rainfall Dipole, is less prominent in Fig. 4.3a.

To investigate the influence of the declustering scheme on the resulting network pattern, we show the node degree of the corresponding network in Fig. 4.3b. In comparison to Fig. 4.3a, the observed features are easier to distinguish and appear to be well pronounced in Fig. 4.3b. Firstly, the low degree band at the position of the ITCZ is narrower and longer extending to the coastline of South America. Secondly, the Amazon basin, where most moisture recycling takes place is marked with elevated degree values which can be differentiated from the rather blurry signature in Fig. 4.3a. Thirdly, the moisture pathway along the East Central Andes is highlighted as well as both exit zones in SESA and SEBRA, which exhibit a comparably high node degree. The differences between the two functional climate networks and their node degree (in Fig. 4.3a and Fig. 4.3b) are directly caused by the implementation of the correction scheme, which prevents the underrepresentation of regions with highly clustered events [77, 78].

We do not only investigate the results of an application of the correction scheme but also aim for carrying out a comparison between the ECA and the ES. Hence, we construct networks utilizing the ECA (for definition see section 2.2.2) and set the according parameters to $\tau = 0$ and $\Delta T = 3$ in order to ensure that we study similar time scales as in the application of the ES (and consider immediate coincidences to coincidences with a lag up to 3 days). In addition, we utilize the average symmetrization (Eq. 2.26) as we are not specifically interested in directed relations between extreme events at different grid points. Other than the method to compute the similarity between the time series we do not change the analysis setup and conduct the analysis first without the correction scheme.

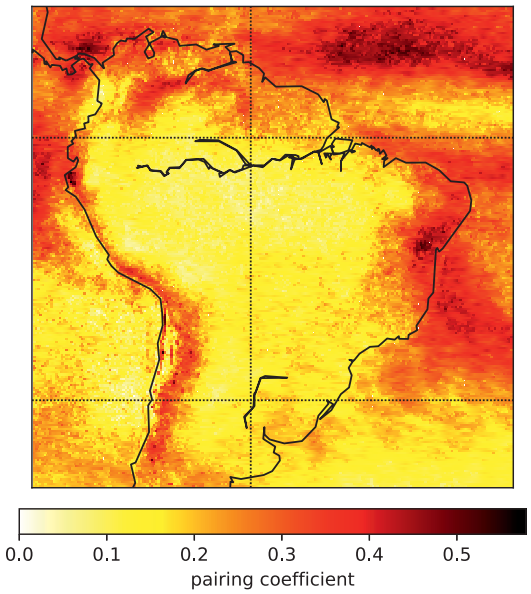

Figure 4.4.: Pairing coefficient of heavy daily precipitation events.

The resulting node degree of the corresponding network is shown in Fig. 4.3c and exhibits different patterns in comparison to the ES-derived networks. The main similarity is the signature of the ITCZ as a band of low degree and an elevated degree values in SEBRA. As the degree maximum north of the ITCZ is characterized by a larger spatial extent and larger degree values and we simultaneously have, as before, limited the edge density to $\rho = 0.02$, other features like the SALLJ or the moisture exit zone in SESA are only partially visible.

With the application of the correction scheme (Fig. 4.3d), we observe a considerable change of the degree pattern and observe very similar features as in the application of the ES with the correction scheme (Fig. 4.3b). The main differences between the node degree of the corrected versions of both methods (Fig. 4.3b and Fig. 4.3d) are the only vague representation of the SALLJ and an increased separation of the two degree peaks related to the South American Rainfall Dipole in the ECA derived network.

4.5.2. Event pairing and degree correlations of differently constructed networks

To further investigate and better understand the differences and similarities between the node degree pattern in Fig. 4.3, we study the so-called pairing coefficient [77]

$$P_i = \frac{1}{s_i - 1} \sum_{l=1}^{s_i - 1} \delta[\left(t_{l+1}^i - t_l^i\right) - 1] \tag{4.1}$$

which quantifies the temporal clustering of events in time series and takes values between $P_i = 0$ (no clustered events) and $P_i = 1$ (all events on subsequent time steps). Note that s_i denotes the number of events in the time series at grid point i and $\delta(\cdot)$ only takes value 1 for vanishing argument. In addition, in this definition we measure time unit-less. As the pairing coefficient is a measure, which we compute for each time series separately, we can in a straightforward manner analyze the previously studied daily event time series (without the application of the correction scheme) and show the resulting values in Fig. 4.4.

Although the pairing coefficient only refers to the information encoded in the single time series, Fig. 4.4 reveals several features of the SAMS and indicates why ES and ECA over- or underestimate the degree in some regions.

In Fig. 4.4 we recognize the ITCZ as a low-value band and especially the region north of the ITCZ as a region of high pairing coefficient. In addition, rainfall events over SEBRA also appear to be highly paired. Comparing Fig. 4.4 with the differences between Fig. 4.3a and Fig. 4.3b or Fig. 4.3c and Fig. 4.3d, respectively, reveals that these two regions are among the most striking differences. Here, the uncorrected ES systematically underrepresents regions with a high pairing coefficient, whereas the application of the ECA leads to an overestimation of the node degree.

Regarding the ES, we infer that the systematic underrepresentation of grid points with high pairing is a consequence of the already explained bias and, thus, results from overly small local coincidence intervals [77]. For the ECA, there does not exist a known generic bias that follows from the definition of event coincidence and we, therefore, suspect that physical interpretation and event definition plays a crucial role and explains the overestimation in the present case. As heavy precipitation events are often related to major weather systems, we expect that in some regions, heavy rainfall events do not only occur over a period of some days but also occur at many neighboring grid points. As the ECA counts all of these events at the subsequent time steps at the different grid points to coincide (see section 2.2.2 for definitions related to the ECA), such large weather systems lead to high coincidence rates and, thereby, an elevated degree. Therefore, the correction scheme, in this case, corrects for the physical misinterpretation of subsequent events as singular events in contrast to one long-lasting persistent extreme event.

In addition, we want to emphasize that neither the degree fields of the different analysis setups nor the pairing coefficient pattern exhibit a direct visual correspondence to the features in Fig. 4.2 and we, therefore, estimate our analysis setup as sufficient to draw the stated conclusions.

Furthermore, we emphasize the differences between the ES and ECA based networks by investigating the degree (Eq. 2.31) of the single nodes and their relation to the pairing coefficient by showing the respective scatter plot in Fig. 4.5. Figure 4.5a features the degree of the nodes in the uncorrected ES based network on the x-axis and the degree of nodes in the uncorrected ECA based network on the y-axis as a scatter plot, color-coded by the respective pairing coefficient. Figure 4.5a supports the previous finding by indicating that only nodes with a high pairing coefficient can achieve a high degree in the uncorrected ECA based network. Vice versa, we observe

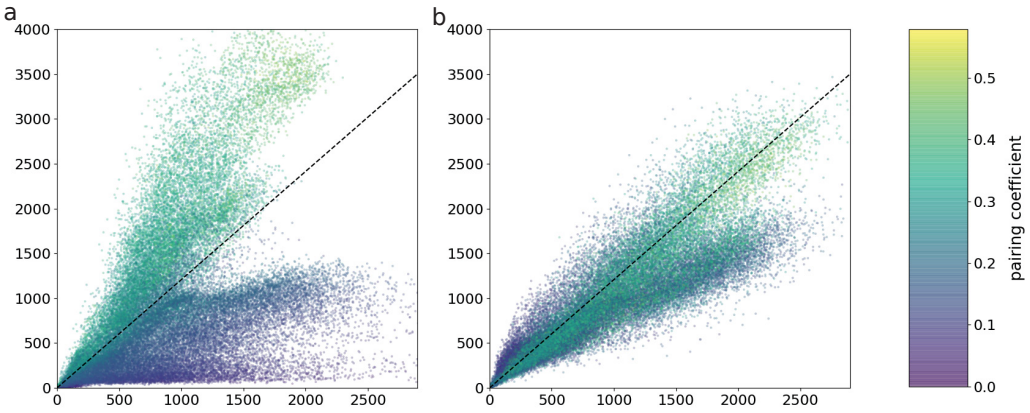

Figure 4.5.: Degree–degree scatter plot for the uncorrected (left) and corrected (right) versions of ES and ECA. The x axes show the degree of a node in the ES based network, whereas the y axes show the degree in the ECA based network. The color code indicates the respective pairing coefficient, the dashed line the line of identity.

that only nodes with low and intermediate pairing coefficient exhibit a high degree in the uncorrected ES based network. These two opposite effects lead to the clear separation of low and high pairing coefficient in Fig. 4.5a.

Figure 4.5b shows the same scatter plot but now for the degree with the utilization of the corrected synchrony measures. In line with the high similarity in Fig. 4.3, the scattered points are now somewhat aligned along the diagonal indicating that a high degree in the corrected ES based network highly correlates with a high degree in the corrected ECA based network. In addition, there is no clear separation between high and low pairing values left after the application of the declustering scheme. Whereas we explain the high similarity by the relative short maximum coincidence windows of 3 days for both methods, which only allows for a few different values of the dynamic coincidence interval of the ES, we assume that the differences are due to the structurally distinct definition of the two methods.

Before we present a second case study highlighting the advantages of both event synchrony measures, we want to emphasize that we present an extension of the method intercomparison in Appendix B.

4.6. Case study 2: Tracking cascading heavy rainfall across scales

In the second case study, we aim for emphasizing the different advantages of ES and ECA to track rainfall cascades. Similar to the previous case, we make use of an already conducted study [5], where the authors have revealed a mechanism leading to rainfall cascades from SESA to the East Central Andes using a corrected ES based directed network approach. To access the time scales, pathways, and direction of the rainfall extremes, the authors combine the network divergence (the difference between

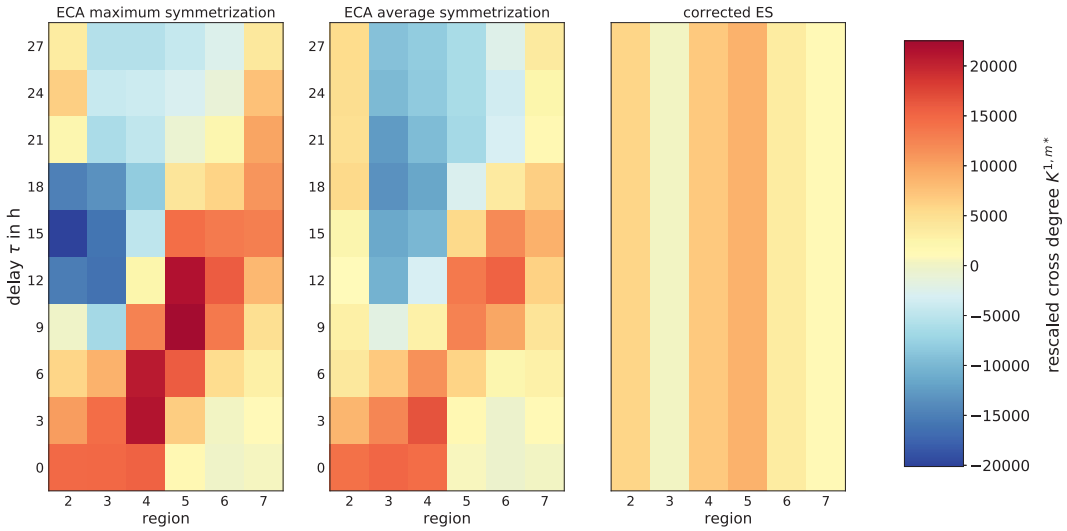

Figure 4.6.: Rescaled total cross degree $K^{1,m*}$ between region 1 and the other boxes in Figure 4.1 for (left) ECA using the maximum of the pairwise event coincidence rates, (center) ECA using the mean of the pairwise rates, and (right) corrected ES, allowing for a varying delay in the ECA based networks.

in- and out-degree in directed networks, see Eq. 2.8) with a comprehensive scheme to track heavy precipitation events occurring successively in 7 regions between the East Central Andes and SESA (indicated by the boxes in Fig. 4.1).

In contrast to this study, we here aim for a purely network-driven approach utilizing the ES and the ECA in multiple different parameter settings. To be able to track the traversing rainfall extremes in detail, we here utilize the 3-hourly TRMM data set and consider rainfall to be extreme if the rainfall exceeds the 98%th percentile. For each pair of the event time series, we first employ the ECA with the average (Eq. 2.26) and the maximum symmetrization (Eq. 2.27) and set the global coincidence interval (Eq. 2.22) to $\Delta T = 1$ while we vary $\tau \in [0, 9]$. Finally, we threshold the resulting similarity matrix $\mathbf{S}$ at some value to obtain a link density of $\rho = 0.05$. These different parameter settings ensure that we stepwise (with increasing τ) track synchronous events in a small time window which at the same time makes an application of the correction scheme obsolete. Second, we apply the cluster-corrected ES without setting parameters (no upper bound for the dynamic coincidence interval (Eq. 2.17) and no time lag) and construct a functional climate network with the same link density.

In each of the described parameter settings, we measure the total cross degree (Eq. 2.9) [30, 56] between different regions (here the 7 regions indicated by the described boxes).

Since the connectivity between spatially disconnected regions depends on their physical distance, we further rescale the total cross degree $k^{1,n}$ between region 1 and the other regions by subtracting the cross degree $k^{1,r;e}$ we would expect from the link distance distribution of the entire network, considering the corresponding fraction of

area the respective region covers in that specific distance range. In particular, we, therefore, calculate

$$k^{1,n*} = k^{1,n} - k^{1,n;e} = (1 - F_n)k^{1,n} \tag{4.2}$$

with

$$F_n = \frac{|E^{1,n}|}{\sum_{e \in E} \left[\Theta\left(d(e) - d^{1,n}_{min}\right) - \Theta\left(d(e) - d^{1,n}_{max}\right)\right]} \tag{4.3}$$

denoting the ratio between the total number of *possible* edges between region 1 and the other region n and the *actual* edges in the network in the specific range of link distances between both regions. Note that $d(e)$ symbolizes distance traversed by the edge $e \in E$ and $\Theta(\bullet)$ the Heaviside step-function.

The resulting rescaled cross degree for the different parameter settings is shown in Fig. 4.6 and highlights the capabilities of the different methods.

First, we emphasize the different results of the rescaled cross degree between region 1 and the other regions for the different delays τ. While we observe elevated rescaled cross degree between region 1 and regions 2,3 and 4 (close to region 1) for small values of τ, we notice a gradual shift towards a maximum of the rescaled cross degree between region 1 and region 6 and 7. This trend is coherent for both symmetrization methods shown in Fig. 4.6a and Fig. 4.6b.

Second, we find larger positive and negative values of the rescaled cross degree and, thus, a more coherent tracking of the rainfall extremes for the maximum symmetrization (Eq. 2.27) shown in Fig. 4.6a. As already implied by definition, this option emphasizes the directionality of event coincidence and better suits the present task of tracking directed rainfall cascades from SESA to the East Central Andes. In addition, the maximum symmetrization leads to negative and, hence, fewer connections than expected between region 1 and $2, 3$ for delays between 9h-18h. In contrast, we observe positive rescaled cross degree between region 1 and 2 for all parameter settings when applying the average symmetrization (Eq. 2.26). We suspect that this is a result of weather systems causing coherent rainfall in SESA but no directed cascade of rainfall extremes from SESA towards the Andes. Although the cascade is better represented in Fig. 4.6a, knowing the temporal resolution of 3h enables us to estimate the time frame of the rainfall cascades to values between 18h-24h based on Fig. 4.6a and Fig. 4.6b which agrees with the earlier finding from Boers et al. [5].

Third, we observe a positive rescaled cross degree between region 1 and the other regions in the corrected ES based network (see Fig. 4.6c). This is a remarkable result as we have not adapted any parameter to tailor our analysis. Although the values are rather low in comparison to the maxima in the ECA based networks and the result does not allow for an estimation of the directionality or time frame, the dynamic nature of the coincidence interval of the ES enables us to capture the rainfall cascade

at once calculating only a single network representation.

4.7. Summary

Event synchrony measures are crucial methods to study climate extremes with the help of Network Science. Thus, it is an integral part of research to investigate possible drawbacks and biases as well as differences between potential approaches. In this chapter, we have shown that Event Synchronization (ES) and Event Coincidence Analysis (ECA) are both prone to differently sourced biases which can be corrected utilizing a declustering scheme.

In a first case study, we have confirmed that ES has a systematic problem of capturing temporally clustered events and, therefore, should always be applied in combination with the correction scheme. The results of the ECA based networks show that preprocessing of the data and a correct interpretation of event clusters as long persistent events can serve as physical reasoning of the declustering scheme and thus correct for corresponding biases.

In a second case study, we have highlighted the different features of the event synchrony measures. Whereas the dynamic character of the ES enables us to grasp event cascades as a whole, ECA helps to reveal the direction and timescales of heavy precipitation events by the possibility to tailor and adapt the corresponding parameters.

Chapter 5.

Spatiotemporal pattern of heavy rainfall during the East Asian Summer Monsoon

5.1. Introduction

After setting the stage for synchrony measures by studying their advantages and possible biases in the previous chapter, we now continue by utilizing event synchrony to disentangle synchronous rainfall during the East Asian Summer Monsoon (EASM).

Heavy rainfall related to the EASM occurs regularly in the period from May to July in South East Asia [108, 109]. The EASM associated rainfall band which is known as *Mei-yu* over China and *Baiu* over Japan gradually moves northward during that period and distributes heavy rainfall over China and the Japanese Archipelago [108, 110, 111]. Studying its temporal evolution as well as the formation and the withdrawal is not only important for agriculture but also crucial to save people's life from natural hazards such as rainfall-induced landslides or floods, which are considerable scenarios based on persistent rainfall in Japan [112]. Whereas a coherent prediction scheme for the monsoon onset in Japan does not exist, it is also up to discussion to which extent the different already identified drivers of the EASM are interdependent.

To contribute some new insights to these open research questions, we investigate heavy precipitation events during the EASM with the help of Event Coincidence Analysis (ECA). In particular, we utilize a sliding window approach to capture the temporal development of eventually synchronous rainfall in the eastern part of the EASM and identify the formation of a double band of coherent heavy precipitation which coincides with the onset of the monsoon season in the respective area. In addition, we study rainfall composites and reveal the interplay of different drivers as responsible for the formation of the double band. With this approach, we tie in with a recent study [75] where the authors have analyzed the synchronization of heavy rainfall in June and July using a static network. In this work, we do not only enlarge the temporal (April to August) and spatial domain ($110°\mathrm{E} - 150°\mathrm{E}, 10°\mathrm{N} - 50°\mathrm{N}$) but focus on the temporal evolution *within* this period.

Accordingly, we first introduce the climatological setting and a static network utilizing data covering the whole period of the EASM. Second, we study the temporal

network evolution before, during, and after the monsoon season. To specifically study the Baiu associated rainfall band, we further investigate the community structure of the underlying temporal network as well as analyze composites related to different periods of EASM emergence. We explain possible atmospheric drivers of the formation of the rainfall band by examining different climatic variables and their temporal change. Finally, we summarize our findings in the last section of this chapter.

The results presented and the figures shown in this chapter are based on publication P_4 (*Wolf, F., Ozturk, Cheung, K., U., Donner, R. V. (2020): Spatiotemporal synchronization patterns of heavy rainfall events during the East Asian Baiu season. Earth Syst. Dynam. Discuss., in review*). We thank Copernicus for the kind permission to reuse/adapt the content and figures.

5.2. The East Asian Summer Monsoon

In vast regions in Asia, annual rainfall is considerably attributed to monsoon systems. From the Arabic peninsula to the Japanese Archipelago regular persistent periods of rainfall can be associated with different branches of the Asian Monsoon System. Whereas the western branch of the Indian Monsoon System (ISM) extends over the Arabic Sea, its eastern counterpart is initiated over the Bay of Bengal and the eastern Indian Ocean [113]. Closely connected to the ISM, the East Asian Summer Monsoon distributes substantial rainfall over China and the Philippines (as Meiyu), the Korea Peninsula (as Changma), and the Japanese Archipelago as (Baiu) [114–116]. As we especially focus on rainfall over the Japanese Archipelago we stick to the term *Baiu* in the following. In the time between May and July, this most eastern branch of the Asian Monsoon System gradually migrates northwards, mostly driven by the Northwest Pacific Subtropical High (NPSH) and the corresponding upper-level jet [111, 117].

Emerging from latitudes around 25°N the Baiu migrates northwards up to latitudes around 40°N in around 40 days. During this period it is characterized as a quasi-stationary subtropical frontal zone which is facilitated with moisture from the tropical Western Pacific, the South China Sea, and the Bay of Bengal. This moisture is transported eastwards by the NPSH associated high-level jet along the front which ranges east-to-northeastward from the Tibetan Plateau to Southern Japan [108, 109]. In this frontal zone, the northeastwards transported warm maritime air masses converge with cool polar air resulting in heavy rainfall events [118]. The further migration of the NPSH and an abrupt northward shift of the subtropical jet together with enhanced convection over the western Pacific end the Baiu season in late July [108, 110, 111]. Especially due to the multiple couplings to other large scale circulation systems or climate variabilities such as the El Niño Southern Oscillation, the Antarctic Oscillation, the ISM and the corresponding South Asian Anticyclone (SAA), the onset and the strength of the Baiu varies interannually and is, thus, only partially predictable [116, 119–121].

5.3. Data, network construction, and parameter choices

As we are interested in heavy rainfall data at high spatial resolution, we employ daily rainfall estimates from the Tropical Rainfall Measurement Mission (TRMM, version 3B42 V7) [25] covering a period between 1998 and 2018 with a spatial resolution of $0.25° \times 0.25°$ latitude and longitude.

For our analysis, we consider rainfall exceeding the 90th percentile as a heavy rainfall event. The choice of this value is not only in line with previous work [75] but also assures a large enough number of events to compute meaningful similarity values with event synchrony measures (3 events per year using a sliding window of 30 days). Setting the event threshold even lower would increase the number of events but contradict the assumption of *rare and heavy* rainfall events. This local threshold (defined for each grid point) also ensures that we achieve the same number of events at each grid point and adaptively change the threshold according to the general rainfall at the different locations. A small number of dry areas with insufficient numbers of rainfall days are mostly located in inner Mongolia and are neglected in the following as they do not cause visible signatures in the network pattern.

To calculate the similarity between the event time series, we utilize Event Coincidence Analysis (ECA) (for details, see section 2.2.2 and [73]). We consider ECA to be an ideal method to capture coinciding rainfall at different locations within an adaptive time window as ECA is not directly susceptible to the clustering bias on which we have elaborated in the previous chapter (chapter 4). As shown, Event Synchronization, an alternative method, should always be used in combination with a correction scheme that removes event clusters [5]. In this study, we already suffer from a very low number of events per time series and do not want to reduce the number by the application of a correction scheme. In addition, we assume that the study area is small enough to assure that we do not observe a possible differing representation of areas with rainfall associated with distinct climatology by the ECA (as shown in chapter 4). This methodological setting also has the advance of constant numbers of events (disregarding dry areas) and, thus, quick computation times which is essential for a sliding window network analysis.

We have set up the analysis framework in line with the analysis in the first case study of the previous chapter (section 4.5) as we assume that similar atmospheric drivers govern the potential synchrony of extreme events. For the ECA, we choose zero time lag ($\tau = 0$) as we are not specifically interested in lagged coincidence (we also want to consider instant or subdaily coincidence of heavy rainfall) and choose $\Delta T = 3$ as the size of the global coincidence interval (Eq. 2.22), as we consider the typical period of directly correlated atmospheric pattern to be 3 or fewer days. In addition, we employ the average symmetrization (Eq. 2.26) if not stated differently as we do mainly not aim for highlighting some kind of directed interrelation.

Finally, we construct the networks based on ECA-derived similarity values by thresholding the similarity matrix (Eq. 2.26) to obtain a link density (Eq. 2.1) of $\rho = 5\%$. This is in agreement with a previous study investigating heavy rainfall extremes using a static network approach [75] and ensures considerable interconnectivity in the

regionally confined study area. To account for biases due to the regional confinement we make use of the framework from Rheinwalt et al. [85] (introduced in chapter 2) to obtain geometry corrected network measures.

In the following, we utilize the described setup and only vary the time window of considered days per year which we accordingly specify.

5.4. Mean EASM network pattern

To tie in with a previous study [75], where the authors have investigated a static network based on heavy rainfall over the Japanese archipelago, we initially study a similar setup but greatly enlarge the temporal and spatial domain. In Fig. 5.1 we illustrate the seasonal mean features of the EASM in the study area.

Figure 5.1a,b illustrate the seasonal mean rainfall distribution between April and July. Besides the heavy rainfall in the southern part of the study area which is related to the northward shifted inter-tropical convergence zone (ITCZ) in boreal summer, we recognize a rainfall band extending from southeastern China over Taiwan and Okinawa to Kyushu, Honshu, and the Korean Peninsula.

In the corresponding network which we base on heavy rainfall events during the 100 days between April 25th and August 3rd and show in Fig. 5.1c,d, we notice different features. We observe elevated node degree (Eq. 2.31) and average link distance (Eq. 2.30) in the southern part which is most likely related to highly synchronous rainfall distributed along the ITCZ. Furthermore, we identify band-like structures in the degree and link distance which are aligned in a Southwest to Northeast direction with two fuzzy peaks, one in the Japanese Sea and another South of Honshu between 20°N and 30°N. In the following, we demonstrate that the evolution of this double band is highly related to the evolution of the Baiu front.

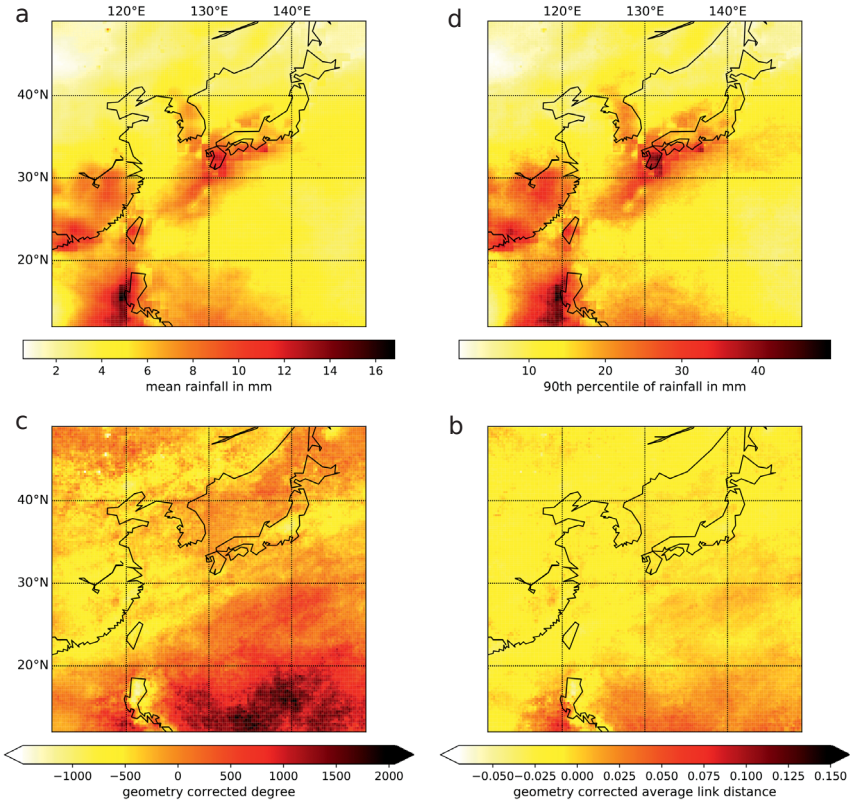

Figure 5.1.: (a) Mean and (b) 90th percentile of daily rainfall sums in the study region between April 25th and August 3rd. (c) Geometry corrected degree and (d) link distance of the climate network based on heavy rainfall events between April 25th and August 3rd.

5.5. Network evolution during the EASM

As stated above, the Baiu onsets in the first weeks of May over Okinawa [111]. To differentiate between the active Baiu and the precursing and subsequent phases, we separately show the networks from rainfall between May and July (active phase) in Fig. 5.2 and from April and August (before and after) in Fig. 5.3.

Figure 5.2 features the node degree (top panels) and link distance (bottom panels) of the networks based on heavy precipitation events in May, June, and July (left to right). We observe that synchronous rainfall during the EASM is closely related to the emergence of a double band structure. In May (Fig. 5.2a,d), the double band is characterized by a region of elevated node degree south of the actual position of the front (and west of Okinawa) and another in the Japanese Sea. In June, especially the northern part of the double band is shifted northwards (which is related to the gradual progression of the Baiu front). Simultaneously, the double band structure appears as clearly marked peaks of elevated degree. The findings are both supported by a corresponding evolution of the average link distance, where the double band is

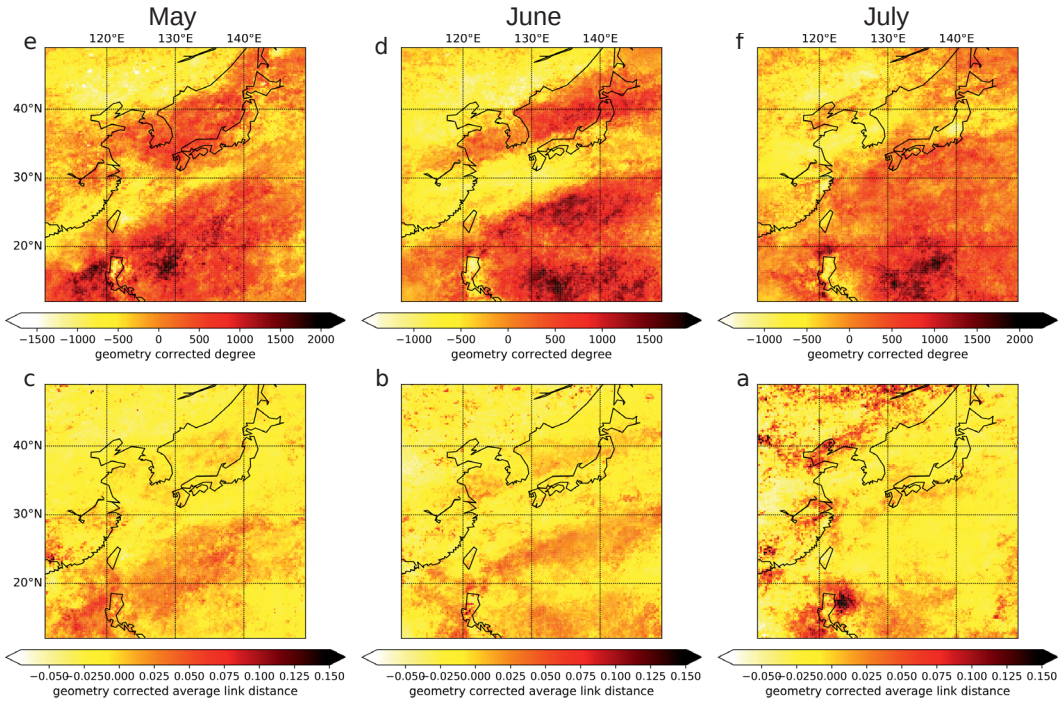

Figure 5.2.: Geometry-corrected degree (a,b,c) and link distance (d,e,f) of networks based on data from May (a,c), June (b,e) and July (c,f).

not as clearly established but shares the described emergence. To the end of the active Baiu phase, the southern band has shifted further northwards and covers the western parts of Honshu, while the northern band has vanished. In addition, we observe a complete change of the corresponding link distance pattern which we assume to be related to the breakdown of the northern band.

To shed light on the formation and breakdown of the double band structure, we additionally show the networks based on heavy rainfall events from April and August in Fig. 5.3.

In April (Fig. 5.3a,c), we already observe elevated degree and link distance values in the Japanese Sea and the Korean Peninsula which coincides with the location of the previously mentioned northern band. Furthermore, we observe a band-like pattern of high degree and link distance in the South of Honshu. In contrast to the double band structure in May and June, these two bands are not well separated. In August (Fig. 5.3b,d) not only the link distance pattern has changed completely but also the southern band in the degree has vanished (in comparison to July Fig. 5.2a,c) and new patterns emerge.

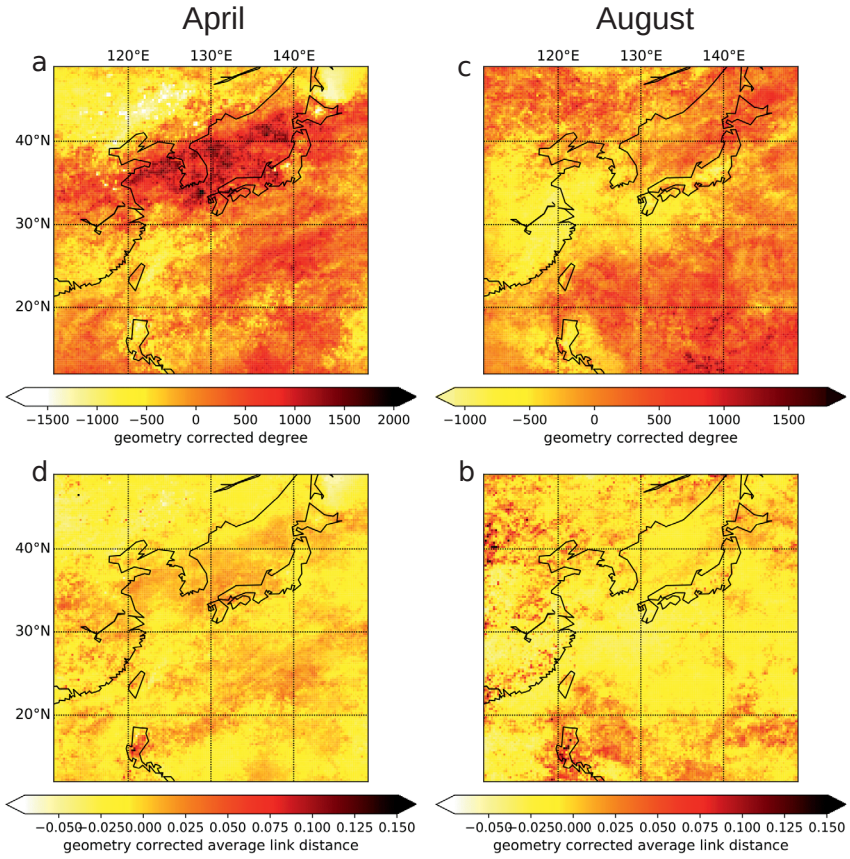

Figure 5.3.: Geometry-corrected degree (a,b) and link distance (c,d) of networks based on data from April (a,c) and August (b,d).

5.6. Temporal interconnectivity of the rainfall band

The previously conducted network analysis has shown that the Baiu phase is characterized by the emergence of a double band structure. As elevated link distance in combination with the high degree in both regions is already an indication of possible interconnectivity between the two regions, we continue by analyzing the community structure (for details see section 2.1.4). We have inferred the networks' communities by utlizing the *Infomap*-algorithm [65] (minimizing the description length, see Eq. 2.16) of the temporal networks to confirm this initial suspicion. In Fig. 5.4a, we show the communities of the network based on the heavy rainfall events occurring in the 30 day window starting on June 14th (ending July 13th). In agreement with the observed double band structure with an orientation from Southwest to Northeast, we find alike arranged and spatially seperated communities. Specifically, we find a spatially seperated community (purple regions in Fig. 5.4a) of which the respective locations coincide with the positions of the two previously mentioned bands. As communities are subsets of nodes in a network with high topological interconnectivity

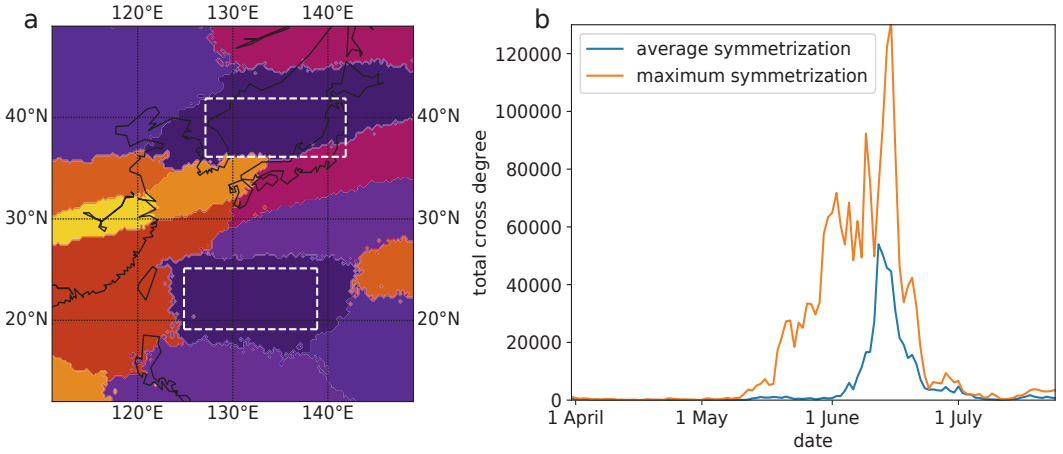

Figure 5.4.: (a) Community structure of the network based on heavy rainfall events between June 14th and July 13th. (b) Total cross degree between the two regions marked in white in panel (a) in the time between April and July.

we hereby confirm that at least in this time window a considerable number of edges connect the two regions of the double band.

To further investigate the temporal evolution of this interconnectivity, we measure the total cross degree (Eq. 2.9) between two boxes, which we locate according to the communities (white boxes in Fig. 5.4a) and, thus, the double band.

For quantifying the mutual linkage of the two regions, we measure the total cross degree which only depends on the edges connecting the regions. To separate potentially different flavors of the directionality of the double band structure, we perform this analysis for both, the average (Eq. 2.26) and the maximum symmetrization (Eq. 2.27) of the ECA and show the results in Fig. 5.4b. In Fig. 5.4b, we observe a marked increase of the total cross degree after May 9th for the maximum symmetrization and a delayed increase of the cross degree for the average symmetrization while both measures peak around mid-June and rapidly decrease after this maximum. Here we want to note that the cross degree is calculated for the 30 days window following the date highlighted in Fig. 5.4b.

We interpret this result as further proof of the mutual connectivity between the bands which is present in both network configurations and coincides with the onset of the Baiu for the average symmetrization. The precursing increase of the cross degree in the maximum symmetrization can either be attributed to the earlier onset of the Baiu in the region around Okinawa or caused by a direct precursor of the bidirectional interrelation. However, the existence of a directional coupling of the heavy rainfall events which transforms into a more bidirectional interrelation is a direct indication of some physical atmospheric process controlling heavy rainfall in spatially distinct regions.

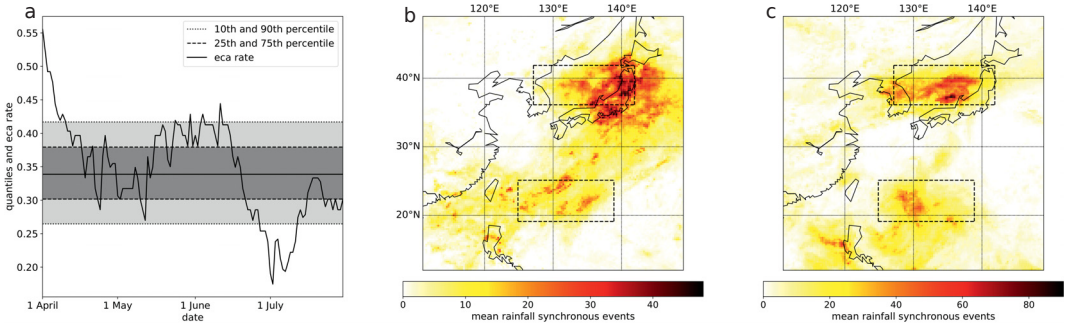

Figure 5.5.: (a) Event coincidence rates between the days with the 10% most synchronous events in the regions marked in Fig. 5.4a for the same sliding windows as in Fig. 5.4(b). The shadings refer to the quantiles (0.1, 0.25, 0.75, 0.9) estimated from 1000 surrogate event time series based on random shuffling.(b),(c) Rainfall composites based on days at which the number of rainfall events exceeds the 90th percentile in both regions between (b) 4 April–3 May and (c) 14 June–13 July, respectively.

5.7. Temporal evolution of the rainfall band

Comparing the locations of the mean rainfall maxima (see Fig. 5.1a.b) and the location of the double band (see white boxes in Fig. 5.4a), we observe that both positions do not directly spatially coincide. This implies that the heavy rainfall events which lead to the formation of the double band in the network measures are not the overall strongest maxima nor do they directly relate to the actual position of the front. If that would be the case then at least one region should cover Kyushu and Honshu where the Baiu front usually distributes the most severe rainfall.

To shed light on the actual rainfall distribution during the phases when the regions of the double band are characterized by synchronous heavy rainfall, we investigate the number of extreme events in the two regions shown in Fig. 5.4a. For each 30 day window of the running window analysis shown in Fig. 5.4b, we calculate the sum of events in each box. Subsequently, we transform the two corresponding time series of every time step into an event time series (by thresholding with respect to the 90th percentile) and compute the event coincidence rates (Eq. 2.23) with the already utilized parameter setting ($\tau = 0$, $\Delta T = 3$). This results in a time series of coincidence rates for the studied period. In Fig. 5.5a, we show the coincidence rates together with the corresponding quantiles which we compute from 1000 randomly shuffled surrogates with preserved number of events and time series length.

Figure 5.5a features the temporal change of a high number of correlated events and exhibits two maxima, one at the beginning of the study period (phase 1) and one around mid-June where the Baiu front is active and the double band is fully developed (phase 2). In addition to these two phases, we observe a phase of significant asynchrony at the beginning of July. This possibly indicates that during some time after the Baiu-related heavy rainfall precipitation in one region goes in hand with suppressed rainfall in the other. We do not investigate this further in this work, but

Table 5.1.: Days with high numbers of grid cells exhibiting heavy precipitation events in the two regions of interest during Phase 1 (4 April–3 May, left) that have been used for defining the composites shown in Fig. 5.5a. The numbers of grid points in the two regions of interest (and therefore the maximum number of grid points with events per day) are 1375 for the northern region and 1344 for the southern region, respectively.

date	#events northern band	#events southern band
14 April 1998	467	789
12 April 1999	657	1069
25 April 2009	862	409
3 May 2012	527	429
6 April 2013	1168	624
14 April 2015	712	573

Table 5.2.: Days with high numbers of grid cells exhibiting heavy precipitation events in the two regions of interest during Phase 2 (14 June–13 July, right) that have been used for defining the composites shown in Fig. 5.5b. The numbers of grid points in the two regions of interest (and therefore the maximum number of grid points with events per day) are 1375 for the northern region and 1344 for the southern region, respectively.

date	#events northern band	#events southern band
9 July 2002	740	717
13 July 2002	865	478
18 June 2004	729	471
22 June 2011	925	452
17 June 2013	664	597
10 July 2013	443	708
11 July 2013	511	764
5 July 2016	629	719

it might be interesting in future studies.

To investigate the climatological difference between the two periods of high event coincidence rate of which only the latter is additionally characterized by a high cross degree (Fig. 5.4c), we show corresponding composites in Fig. 5.5b,c. Both composites are only based on instantaneous coincidences of over 90th percentile event sums in both regions and feature the mean rainfall distribution in phase 1 and phase 2. Although the composites are based on 6 events (phase 1) and 8 events (phase 2) only and refer to the 30 day time windows after April 4th and June 14th, respectively, they are representative for other time windows in both periods (not shown). We show the details of the event numbers and the respective days in table 5.2.

During the first phase of elevated event coincidence rate, we observe a large region of coherent rainfall covering not only the two regions but also the region in between (Fig. 5.4b). We attribute this pattern to a major weather system distributing rainfall over a large spatial extent. In contrast to this configuration, we observe spatially separated heavy rainfall which is mostly located within both boxes during phase 2 (Fig. 5.4c). This pattern only leads to the high cross degree indicating the interconnectivity between the two regions, as heavy rainfall events most prominently are located within the double band.

5.8. Discussion

All results presented above are not only based on rainfall data but also derived by the application of the network construction scheme and, thus, do not directly relate to specific atmospheric processes. Accordingly, we now investigate composites based on multiple climatic variables (advection, relative vorticity, winds in different altitudes, geopotential height). The discussion is structured as follows: initially, we present the most important drivers which we identify in our climatological analysis and introduce them by reviewing the related literature. In particular, we show monthly means of relative vorticity, 850 and 200 hPa wind and 500 hPa geopotential height (in 5.6) based on the NCEP-DOE-Reanalysis (versions 1 and 2 [122, 123]. Subsequently, we perform two case studies, one based on composites from dates with high inter-band activity (see the previous section) and another based on particularly strong Baiu events to illustrate the importance of the presented drivers.

5.8.1. The South Asian Antocyclone and the North Pacific Subtropical High

The Baiu, as one part of the EASM, is linked to different climatological features in other regions of the EASM. Especially, a close connection between the Baiu and Meiyu (the frontal rainfall over southern China) has been confirmed. During the EASM, south-westerlies transporting moisture towards the Baiu region have been proven to contribute to heavy rainfall over Japan [124]. In addition to these intra-EASM connections, studies have demonstrated that the Indian Summer Monsoon (ISM) is also correlated with the EASM. While the onset of the ISM is closely related to

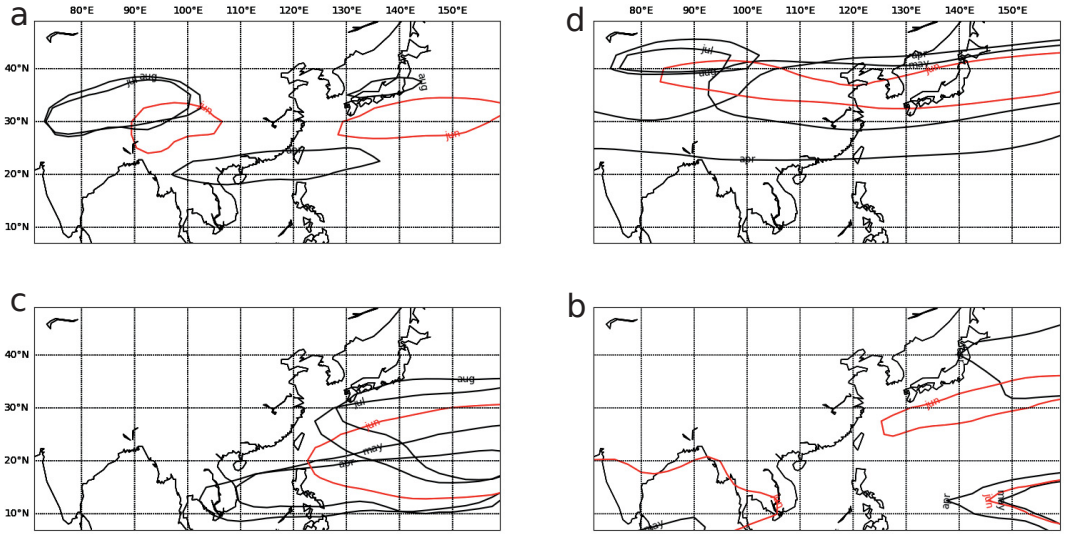

Figure 5.6.: Monthly mean contour of (a) 200 hPa relative vorticity at $2.8 \cdot 10^{-5}$ s^{-1}, (b) 200 hPa isotachs of 30 ms^{-1}, (c) 500 hPa geopotential height at 5880 m, and (d) 850 hPa isotachs of 7 ms^{-1} during April-August. Contours in June are colored red.

the formation of the Baiu and Meiyu frontal systems, the precipitation in Kerala is significantly correlated with Baiu related rainfall at intraseasonal and interannual timescales. This coupling has been explained by the so-called southern mode of teleconnection [109, 116, 125].

Rainfall over Japan during the EASM is particularly influenced by the South Asian Anticyclone (SAA) the intensity of which is correlated with the magnitude of the high-altitude jet connecting the Tibetian plateau with the study region [116]. In addition, the SAA is part of the cyclone-anticyclone pattern of the southern mode of teleconnection emerging from a zonally oriented upper-level wavetrain [109, 125]. To illustrate the development of the SAA between April and August, we show the corresponding relative vorticity at $2.8 \cdot 10^{-5}s^{-1}$ in Fig. 5.6a. While the high altitude vorticity remains low in April and May, the SAA forms in June and persists during July and August. In line with the formation of the SAA in June, we observe a narrow zonal band of strong upper-level wind at about 35°N to 40°N (see Fig. 5.6b) which is an indication of the enhanced high altitude jet. The jet which retreats to the west in the following month is one of the factors which finally drive the formation of the Baiu front. In addition to establishing the high altitude jet, previous work has also shown that a strong SAA is correlated with an enhanced meridional upper-level temperature gradient [116], which may serve as a source for the temperature advection in early June [111, 126] and the consecutive migration of the Baiu front.

The second major driver of the Baiu is the North Pacific Subtropical High (NPSH) which is accompanied by the corresponding low at high altitudes. This low induces a strong high-level jet which reaches down to lower levels at the northern ridge of the NPSH in the latitudes just South of the Japanese Archipelago, and, therefore, serves

as one driver for the Baiu migration [125]. This stable weather system is previously considered as independent of the SAA (correlation coefficient of 0.02 estimated by Li et al. [116]). The comparably persistent high forms in April just north of the tropics and migrates northward during the study period (Fig. 5.6c), in line with the mean Baiu development. In June, the northern ridge of the NPSH is located just south of Japan at 30°N (Fig. 5.6c). As mentioned above, the corresponding westerlies are, thus, also centered at this latitude and reach from Taiwan to longitudes ≥ 160°E. In addition, the low-level winds form a narrow corridor and serve as a major driver of the Baiu onset in this period (Fig. 5.6d). The distinct combination of strong and narrow westerlies just at the northern ridge becomes visible by comparing the pattern from June to the pre- and post-Baiu phases (Fig. 5.6c,d).

In summary, we assume, that the observed double band of synchronous rainfall which is most prominent in the network of June (Fig. 5.2b,e) is a signature of the interplay between the SAA and the NPSH. In contrast to the previous finding of the independence between these two, we suggest that the synchronous heavy precipitation events are possibly triggered and controlled by both and, thus, by their interaction.

In the following, we provide two different approaches to confirm this suspicion by first extending the study from section 5.7 and investigating corresponding composites and second by conducting two case studies of strong Baiu phases.

5.8.2. Climatology during high inter-band activity

To tie in with the network analysis and the identified double band, we investigate composites during the days of high inter-band activity (see section 5.7 and table 5.2 which are most likely responsible for elevated values of cross-degree in the temporal network analysis.

Figure 5.7a,b features the NPSH and the SAA-related jet. In Fig. 5.7a the low values of relative vorticity (blue colors) indicate the position of the NPSH while the clockwise rotating winds transport the moisture towards the southern band of enhanced rainfall. A further indication of the NPSH are the elevated values of geopotential height just south of Japan in Fig. 5.7b. The SAA-related jet which is illustrated by the white contours in Fig. 5.7b serves as the main source for the heavy rainfall in the northern band.

In line with the spatially separated rainfall, we observe two distinct regions of updraft and corresponding enhanced vertical velocity referring to the two bands in Fig. 5.7c. In addition, we show the specific humidity as blue contours which confirm the role of the upper-level jet as a source of moisture for the northern band. Finally, we observe that it is indeed a spatially separated formation of clouds and, thus, two separate weather systems serving the northern and southern band as indicated by the OLR anomaly in Fig.5.7d. Here, the southern band is characterized by very strong anomalies revealing the existence of deep (tropical) convection in the southern and weaker (extratropical induced) convection in the northern band (maximal values not shown by the centered color bar with upper and lower limit).

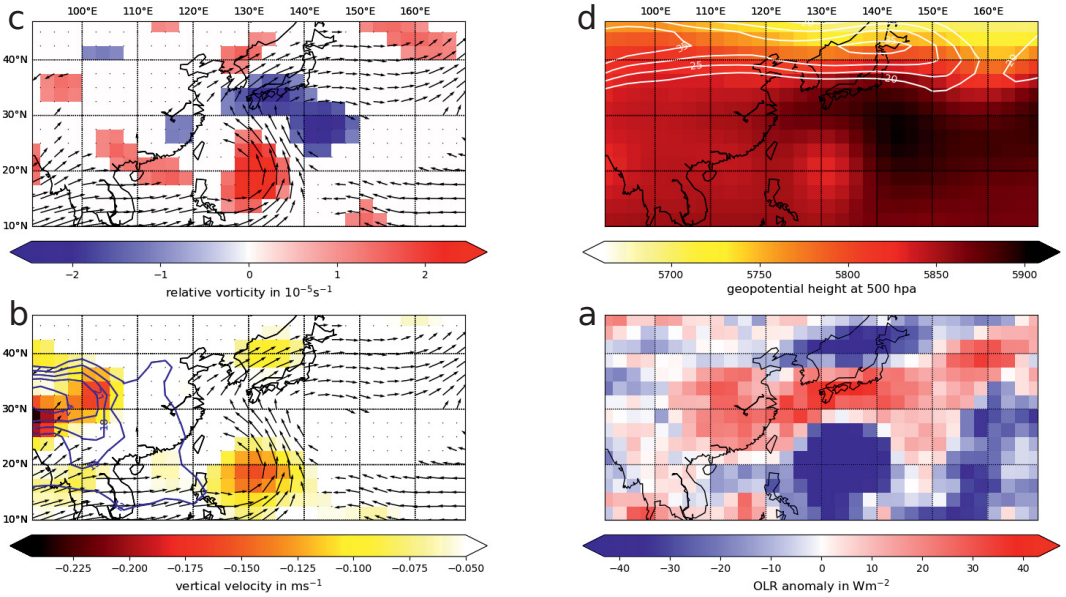

Figure 5.7.: Composites based on the dates with strong inter-band activity (see Fig. 5.5c and table 5.2). (a) Average 850 hPa winds with speed larger than 5 ms^{-1} and relative vorticity larger than 10^{-5} s^{-1} (red tones) and smaller than -10^{-5} s^{-1} (blue tones), (b) Average 500 hPa geopotential height (shaded) and 200 hPa wind speed larger than 20 ms^{-1} (white contours). (c) Average 850 hPa winds with speed larger than 5 ms^{-1}, specific humidity larger than 12 g kgs^{-1} (blue contour, interval 3 g kg^{-1}) and 700 hPa updraft vertical velocity larger than 0.05 Pa s^{-1} (shaded). (d) OLR anomalies with respect to the NOAA long term mean between 1981-2010.

5.8.3. Case studies: periods between 29 June – 3 July 1998 and 25 June – 30 June 1999

To go beyond the identified dates of strong inter-band activity and discuss the heavy rainfall and the related atmospheric processes in other temporal domains we present two case studies from Guan et al. [127] where the authors identified the pair of a cyclone and an anticyclone as the main driver of the frontal development during the active Baiu phase. The authors mention the periods between June 29th and July 3rd, 1998, and June 25th and June 30th, 1999 which coincide with heavy Baiu induced rainfall. Both years (1998 and 1999) are among the rather strong Baiu years. In the first period, rainfall was distributed over the Bay of Bengal and northeastern China, while the latter was characterized by an extended rainfall band ranging from eastern China to Honshu and, therefore, leading to heavy rainfall in Japan (Fig. 5.8a,b). In both cases, the authors attributed the occurrence of heavy rainfall to the existence of a northern shifted anticyclone which they refer to as the Midlatitude Anomalous Cyclone (MAC) and an anticyclone which they term as the Tropical Anomalous Anticyclone (TAAC).

In the period between June 29th and July 3rd, 1998 [127] the distinctive MAC/TAAC

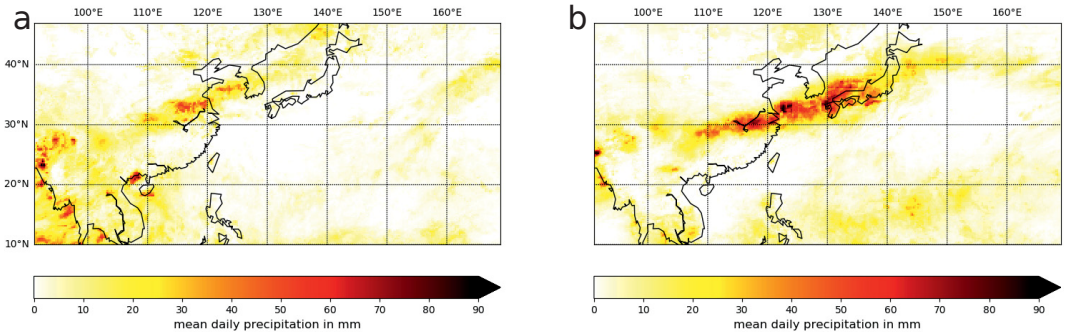

Figure 5.8.: Mean rainfall (a) between 29 June and 3 July 1998 and (b) between 25 and 30 June 1999.

pair intensified the Baiu rainfall and is reflected in the relative vorticity (Fig. 5.9a). The corresponding south-westerlies (arrows in Fig. 5.9a) indicate the cyclonic flow towards the front whereas the gradient in geopotential height confirms the updraft and the corresponding precipitation at the front (Fig. 5.9b). Furthermore, we observe a band of strong upper-level westerlies enhancing the convection at the position of the front (white contours in Fig. 5.9b).

In the latter period, the synoptic factors were organized differently but nevertheless strongly connected to the MAC/TAAC pair. The MAC was significantly weaker while the TAAC appeared to be well developed and facilitated strong south-westerlies. With the absence of the upper-level jet related to the MAC, moisture was transported by these south-westerlies originating from southern Asia (Fig. 5.9c). Another indication for southern Asia as the main moisture source is the strong advection in that area shown in Fig. 5.9d. In addition, enhanced updraft (Fig. 5.9c) which is needed for strong convection [128] organized the latter along the front and initiated the rainfall shown in Fig. 5.8b.

As illustrated by both case studies, the TAAC, which originates from the NPSH is a key factor driving the emergence of the Baiu front. In addition, we have also shown that the upper-level jet (Fig. 5.9b), which most likely arises due to the formation of the SAA facilitates convection at the front. While in both cases the TAAC suppressed rainfall in that region, we strongly assume that the TAAC is a key factor in controlling the rainfall in the southern band which we have identified in the network analysis.

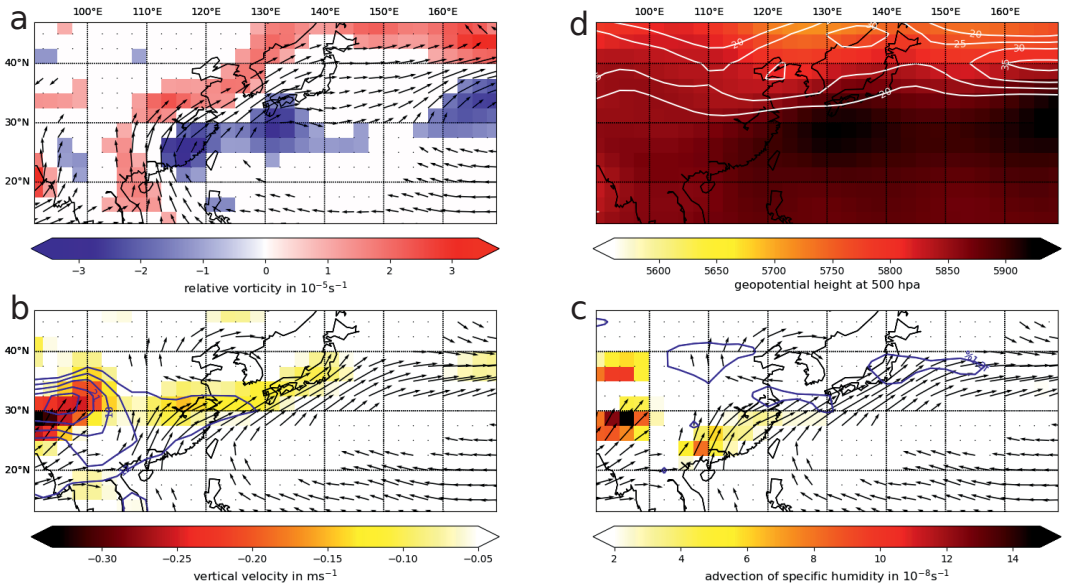

Figure 5.9.: (a) Average (29 June–3 July 1998) 850 hPa winds with speed larger than 5 ms^{-1} and relative vorticity larger than 10^{-5} s^{-1} (red tones) and smaller than -10^{-5} s^{-1} (blue tones), (b) Average (29 June–3 July 1998) 500 hPa geopotential height (shaded) and 200 hPa wind speed larger than 20 ms^{-1} (white contours). (c) Average (25–30 June 1999) 850 hPa winds with speed larger than 5 ms^{-1}, specific humidity larger than 12 g kgs^{-1} (blue contour, interval 3 g kg^{-1}) and 700 hPa updraft vertical velocity larger than 0.05 Pa s^{-1} (shaded). (d) Winds as in panel (c), but combined with specific humidity advection larger than 2×10^{-8} s^{-1} (shaded) and divergence at -0.01 ms^{-1} (i.e., convergence, blue contour).

5.9. Summary

In this chapter, we have analyzed the temporal organization of heavy rainfall during the EASM. In particular, we have investigated the synchrony of heavy rainfall events between April and August utilizing event coincidence analysis and corresponding temporal networks based on a sliding window approach.

In our analysis, we first investigated the monthly evolution of the EASM related network pattern by shifting a corresponding 30 day time window and studying the monthly network pattern. This can be considered as an extension of a previous study [75] as we do not only enlarge the spatial extent of this study but also analyze the temporal evolution of synchronous heavy rainfall events in contrast to the former static approach. In the monthly pattern, we have identified a double band of spatially separated, coherent rainfall. The emergence of this double band, which we have not only observed in degree and link distance pattern but also confirmed by the analysis of the networks' community structure, is closely related to the onset of the Baiu. This has been shown by studying the total cross degree between the two regions of synchronous rainfall by using a sliding window of 30 days which we consecutively shift by 1 day. The corresponding results reveal the onset of a unidirectional relationship

between the two regions in the network after the onset of the Baiu in the Okinawa region and a subsequent bidirectional relation after the onset of the Baiu over Honshu and Kyushu.

By studying the corresponding maxima of event numbers in the two regions of the double band, we have shown that there exist two distinct periods which are characterized by large numbers of coinciding events in both bands, one at the beginning of April and one after mid-June. Whereas the former is related to large scale weather events that distribute rain simultaneously in both bands and the region in between, the latter is characterized by two spatially separated regions of rainfall which coincide with the double band.

By conducting three case studies and investigating corresponding composites based on different variables, we have shown that the double band of synchronous rainfall which to our best knowledge has not been studied before, is closely related to the upper-level jet emerging from the South Asian Anticyclone and low-level south-westerlies originating from the North Pacific Subtropical High. Our results from the network analysis suggest that these two do not behave like independent actors influencing heavy rainfall during the active Baiu phase as previously described in the literature but are somewhat linked via an external variable or mutually dependent drivers. In addition, we have shown that the interdependence between the two bands is temporally coinciding with the Baiu onset, which might set the stage for a new approach for a Baiu onset prediction scheme.

Nevertheless, we stress that the described rainfall pattern is not the prominent pattern during the active Baiu phase, where rainfall is normally distributed along a west-to-east orientated front which gradually migrates northward. Therefore, we assume that the illustrated mechanism possibly only occurs intermittently in the context of some specific large-scale synoptic situations. We outline the corresponding investigations on related questions as a subject of future studies.

Chapter 6.

ITCZ dynamics as seen by network theory

6.1. Introduction

Unlike in the previous chapters, we do not examine extreme events associated with a monsoon system to construct functional climate networks in this chapter. Here, we aim for disentagling the relationship between monthly sea surface temperature variability and tropical rainfall at the inter-tropical convergence zone (ITCZ).

The ITCZ is a band of low pressure emerging from the convergence of the near-surface winds of the Hadley circulation. Driven by the moisture influx of these near-surface winds, regular heavy rainfall affects people and nature in this zone [129]. In total, one-third of the global precipitation falls within the region between 10°N and 10°S [130].

During each year, the position of the ITCZ changes with the position of the Earth on its orbit around the sun. Mainly driven by the hemispherically asymmetric heating, the ITCZ is shifted into the heated hemisphere. Recent work has explained this either by cross-equatorial heat transport which adjusts the energy imbalance [131, 132] or by sea surface temperature-(SST-)induced changes in near-surface moist static energy as well as boundary-layer convergence [133, 134]. Both pictures are well-established in the field and have led to a coherent picture of the time mean ITCZ location. In particular, the time-mean ITCZ position is commonly linked with the time-mean energy transport and the time-mean tropical SST gradient.

In contradiction to these established frameworks, recent model-based works have not only shown that the link between cross-equatorial heat transport or tropical SST-gradient can be not as strong as expected [135], but also climate models commonly fail to correctly reproduce the ITCZ with its characteristic dynamics. Furthermore, state-of-the-art climate models exhibit only rough agreement on the response of climate change due to rising carbon dioxide levels and the corresponding shifts in the annual migration of the ITCZ [23, 136, 137]. To contribute to the increasing effort towards studying idealized models to understand small-scale climate variability and its link to ITCZ dynamics [131, 132, 138, 139], we, here, conduct an alternative approach. In this chapter, we perform a network-based analysis of SST variability by investigating idealized model output from 14 global circulation models (GCMs) where an aquaplanet is studied. In our work, we use this special class of climate models, as we

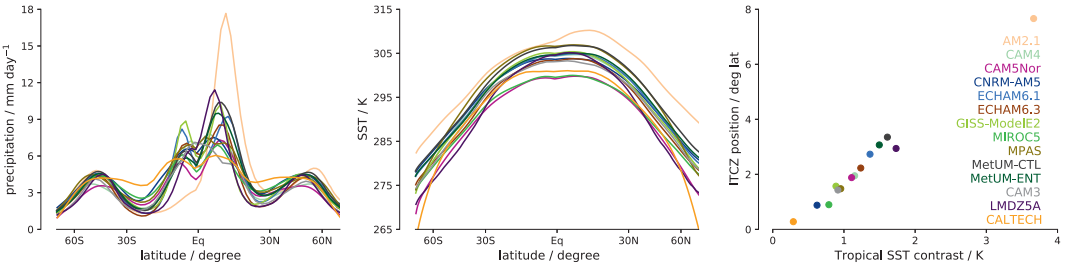

Figure 6.1.: Time- and zonal-mean precipitation (a) and SST (b) in the AquaControl simulations. (c) Correlation between time-mean ITCZ and time-mean tropical SST contrast.

are specifically interested in the unperturbed dynamics of the ITCZ. Neglecting land masses and orography inhibits monsoon-like precipitation anomalies and convection gradients. Therefore, aquaplanets are ideal setups to study the tropical rain belts [86]. In each of the different aquaplanet setups, we investigate the monthly SST pattern and the corresponding zonal mean network measures. In particular, we utilize hierarchical clustering based on the zonal mean node degree and zonal mean average link distance to cluster the 14 GCMs into groups. Finally, we analyze the mean ITCZ positions and SST gradients of the models in the identified clusters and evaluate whether the respective clusters share distinct time-mean ITCZ characteristics. By systematically manipulating connectivity patterns, we iteratively study the importance of extratropical, tropical, and extratropical-tropical connections, regarding their contribution to distinct signatures in the network and the corresponding influence on the clustering output.

Hence, the chapter is organized as follows: we first present the TRACMIP model ensemble which serves as the data source for the study. Subsequently, we introduce the methodological setup and, thereby, present the network construction and hierarchical clustering approach. Thereafter, we show the results of different network analyses. Specifically, we study global networks as well as specific subnetworks to disentangle the importance of different connection patterns. Finally, we discuss and summarize the most striking findings.

The results presented and the figures shown in this chapter are based on publication P_5 (*Wolf, F., Voigt, A., Donner, R. V. (2020): A climate network perspective of the intertropical convergence zone. Earth Syst. Dynam. Discuss., in review*). We thank Copernicus for the kind permission to reuse/adapt the content and figures.

6.2. The TRACMIP model ensemble

The TRACMIP (Tropical Rain belts with an Annual Cycle and a Continent Model Intercomparison Project) is a model ensemble containing simulations of 14 different GCMs which was designed to study central features of the ITCZ. All models have been simulated in two different setups, a first containing an idealized aquaplanet only and a second including an additional idealized continent. Thereby, the TRACMIP allows

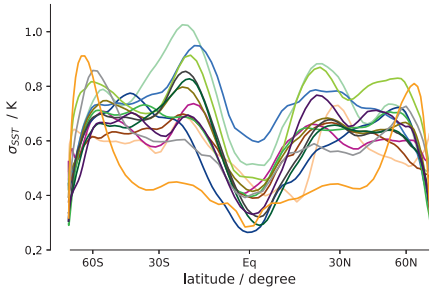

Figure 6.2.: Zonal-mean of the standard deviation of monthly SST time series in the AquaControl simulations. The different models are color-coded as in Fig. 6.1.

for studying the relation between the tropical SST gradient (or the cross-equatorial heat transport) and the ITCZ variability [135]. The model output is available on a regular spherical grid with varying spatial resolution of $1° - 3°$ degrees latitude and longitude. For our analysis, we will solely focus on the SST data. In a previous study [86], the authors have presented a comprehensive summary of the features of the different GCMs and the corresponding ITCZ dynamics.

As we are not specifically interested in differences between the ITCZ migration over land and water masses, we utilize the aquaplanet model runs in this work. These aquaplanet setups include a slab ocean with implemented present-day ocean heat transport and ocean-atmosphere coupling which allows for interactive sea-surface temperatures. Besides, the annual insolation follows the present-day solar insolation and, thus, induces the typical migration of the ITCZ in all models which is located in the northern hemisphere in the time- and zonal-mean.

In this work, we stick to the definition of locating the ITCZ through the precipitation centroid [140], which has also been used by other authors studying the TRACMIP model output [135]. In Fig. 6.1a we show the zonal mean precipitation of the different models to demonstrate the precipitation pattern and the locations of the model-dependent maxima. The corresponding time mean ITCZ positions vary little from the time mean ITCZ locations stated in the already mentioned comprehensive TRACMIP summary [86] but the results which we present in the following are valid for both definitions of the ITCZ location calculation frameworks. To illustrate the close relationship between the SST gradient (which we define by the difference of the SST means between 0°N and 20°N and 0°S and 20°S, respectively; for zonal-mean SST see Fig. 6.1b) and the time- and zonal-mean ITCZ position, we show the corresponding values in Fig. 6.1c.

As already mentioned, previous work [133, 134] has confirmed the close relationship between the time-mean SST contrast and the ITCZ position which we show in Fig. 6.1c. This mechanism has, therefore, been incorporated in many Earth system models. Here, we conduct an alternative approach by evaluating the SST variability of the different aquaplanet setups utilizing functional climate networks. To obtain a basic understanding of the SST variability, we show the zonal mean SST variance of the 14 GCMs in Fig. 6.2 and observe local minima at the poles and in the tropics

and rather constant elevated values in between.

In addition to this present-day scenario, which is called *AquaControl* in TRACMIP and this study, TRACMIP also contains climate change scenarios with quadrupled carbon dioxide concentration (1392 ppmv in comparison to 348 ppmv in AquaControl) to which we refer as *Aqua4xCO2* model runs. This coherent increase in CO_2 results in a global mean temperature increase between $3 - 10$K and slight southward to strong northward shifts (up to 7° latitude) of the time mean ITCZ among the models. Note that Fig. 6.1 and Fig. 6.2 are based on the AquaControl model runs.

6.3. Methodological setup

6.3.1. Network construction and network measures

In this study, we use monthly SST anomalies to construct functional climate networks. Accordingly, we follow the strategy outlined in chapter 2 and employ Pearson correlation (Eq. 2.28) to distill meaningful correlation values to abstract SST pattern into a network representation. Here, we obtain the different adjacency matrices by thresholding the respective Pearson correlation based similarity matrices (Eq. 2.29) by some value such that the resulting link density (Eq.2.1) equals $\rho = 0.005$. This is of the order commonly chosen in global climate networks with high spatial resolution [30, 57].

To avoid biases of the network measures due to the initial conditions of the model runs, we only utilize the in-equilibrium phase of each simulation. Therefore, we only consider the last 30 years of each AquaControl model run and the last 25 years of each Aqua4xCO2 model run for our analysis. The difference in spin-up time arises from a delayed convergence of the Aqua4xCO2 models. As we want to maximize the number of annual cycles studied, we stick to the stated choice. This assures both, completed spin-off of all GCMs and identical time periods for the network construction for the respective model class.

The 14 GCMs within the TRACMIP do not share the same spatial resolution. This implies that keeping the link density constant for all networks directly induces different numbers of links in the different networks. We prefer that over the option of maintaining some constant number of links, as different link densities have been proven to correlate with substantial differences in the network topology [57, 141]. In the following, we adapt our analysis framework to minimize the resulting systematic differences in the network measures.

Although we have studied many different local and global network characteristics (such as local clustering coefficient, transitivity, closeness, etc.), we, here, stick to the analysis of the node degree (Eq. 2.31) and the average link distance (Eq. 2.30). This allows for including topological as well as spatial information of the networks in the analysis and assures the interpretability of the results. In addition, other network measures have only little contributed to other insights and have, therefore, been excluded from the study.

6.3.2. Hierarchical clustering

Motivated by the zonally uniform boundary conditions of the aquaplanet setup, we only analyze the zonal mean network measures in this study. As we are interested in identifying groups of similar SST anomaly patterns, we intend to cluster the different models employing the zonal mean network measures. Here, we utilize hierarchical clustering to achieve this goal.

Initially, we resample the zonal mean network measure distributions to the lowest latitudinal resolution (2.8° latitude, Caltech model) using linear interpolation. Subsequently, we compute the similarity between the resampled zonal mean network measures by utilizing Pearson correlation (Eq. 2.28). This approach ensures that the systematic offset between the zonal mean network measures which is induced by the different spatial resolutions does not have any influence on the similarity, as Pearson correlation is not sensitive to differences resulting from constant factors. We sum the resulting individual correlation scores which we obtain from measuring the similarity between the zonal mean link distance and zonal mean degree for each model pair. Finally, we insert the summed similarity scores which are bounded by $[-2, 2]$ (sum of two values bounded by $[-1, 1]$) in the matrix $\boldsymbol{S}$. Therefore, $\boldsymbol{S}$ has a dimension of 14×14. To cope with the condition of hierarchical clustering methods of input values bounded by $[0, 1]$, we rescale the similarity scores in $\boldsymbol{S}$ by computing

$$s_{ij}^{new} = \frac{s_{ij} - \min \boldsymbol{S}}{\max_{ij}(s_{ij} - \min \boldsymbol{S})}, \tag{6.1}$$

where $\boldsymbol{S}_{new}$ represents the rescaled correlation matrix which we use as an input for the hierarchical clustering method. Note that $\min \boldsymbol{S}$ can be a negative value.

To the end, we cluster the 14 GCMs using the hierarchical clustering method from the python package *scipy* with the default setting of single linkage. This choice is motivated by the central goal to group the most similar models into the same cluster. Single linkage possibly leads to including some outliers into a cluster, which is not a problem due to the relatively low number of models.

In hierarchical clustering methods, entities (here the models) are pairwise merged into clusters based on the input similarity. Due to the stepwise merging of the different models, clusters grow in size and finally end up as one cluster. This procedure is commonly visualized in a dendrogram which is read from left to right. Horizontal lines represent the cophenetic distance (level of similarity between [groups of] models). Vertical lines indicate the merger of some (groups of) models. To obtain the clusters, we can cut the dendrogram at some value of cophenetic distance. All models that have merged into one group below that value are considered as one cluster.

Finally, we acknowledge that there are several different options for clustering models by the means of their zonal network measures. Here, we have chosen this option motivated by the objective to minimize the effect of the different spatial grid resolutions and to maximize the simplicity of the analysis framework (by choosing basic correlation measures).

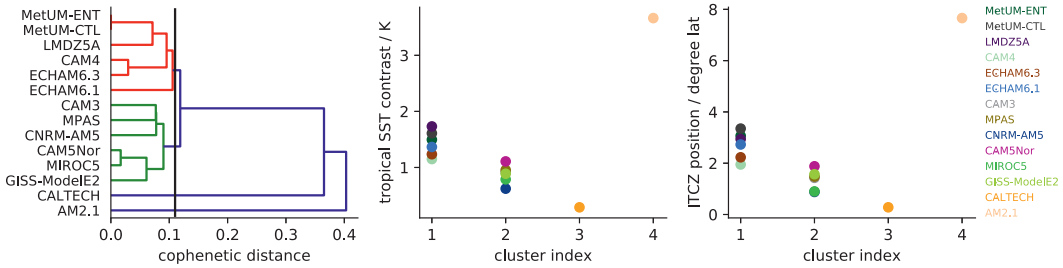

Figure 6.3.: (a) Dendrogram based on the clustering of the zonal-mean network measures utilizing global connection patterns. Models are split into the cluster at the cophenetic distance indicated by the vertical line. Cluster-wise time-mean (b) tropical SST-contrast and (c) ITCZ position.

6.4. Network analysis - global SST correlation pattern

We start by analyzing the global network of the AquaControl model runs. The application of the above-explained framework leads to the dendrogram we show in Fig. 6.3a. Here, we identify four clusters, where the first two consist of six models and the last two are single model clusters. As explained, these results are based on the zonal mean network measures, which we show in Fig. 6.4.

Before we study the zonal network distributions in greater detail, we highlight the clustering regarding the tropical SST contrast and the mean ITCZ position, which we illustrate in Fig. 6.3b,c. Both figures underline the success of the clustering as the models separate well in both measures. In particular, models in cluster 1 exhibit a greater SST contrast and a more northward shifted ITCZ than models in cluster 2. Furthermore, the single model cluster 3 shows almost no tropical SST contrast and an ITCZ close to the equator, whereas the single model cluster 4 is characterized by values even larger than the models in cluster 1. We emphasize that there is no overlap between the clusters regarding the corresponding ranges of the mean ITCZ positions and SST contrasts. As expected from this results, the models also separate by the means of their Southern Hemisphere Hadley circulation strength (minimum of mass stream function: cluster 1: 130-158; cluster 2: 102-129; cluster 3: 53; cluster 4: 266; all values in units of 10^9 kg/s) which is closely related to the mean ITCZ position [132].

The clusters do not only differ regarding their SST contrast and mean ITCZ position but exhibit distinct zonal mean network measure distributions which we show in Fig. 6.4. Figure 6.4 features the normalized (sum of all values along the latitude range equals 1) zonal mean degree (Eq. 2.31, left column) and the zonal mean average link distance (Eq. 2.30, right column) for the different models. In addition, we indicate the ITCZ position of the models in each cluster by the vertical lines.

The models in cluster 1 (top row in Fig. 6.4) exhibit a global maximum at the center of the southern Hadley cell and a coherent minimum of both network measures around the position of the ITCZ. In a study analyzing a single model, which is part of this first cluster (ECHAM6.1) (P_1, see chapter 3) this has been confirmed and

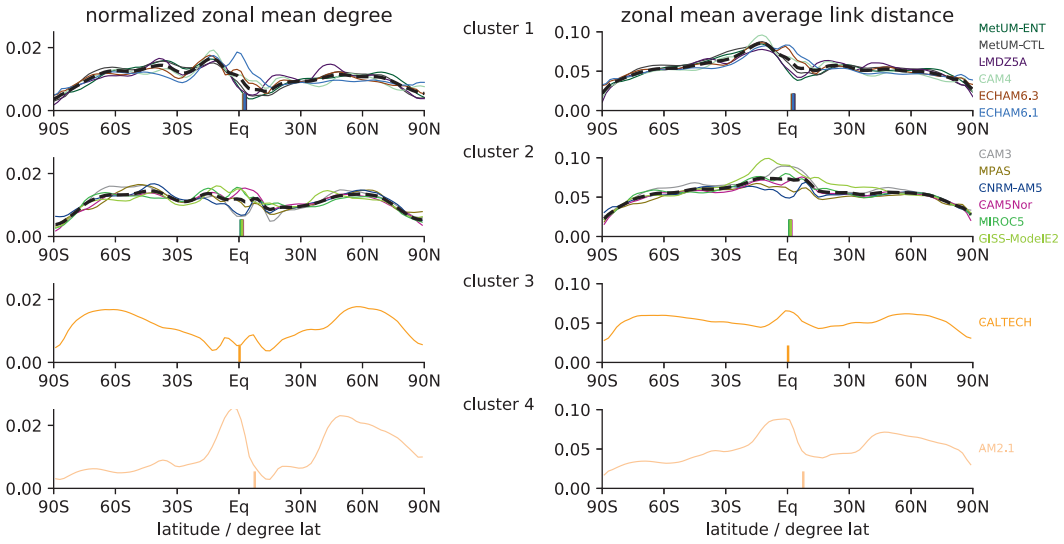

Figure 6.4.: Zonal mean average node degree (left panels) and zonal mean average link distance (right panels) for all four previously identified clusters (vertically ordered, see Fig. 6.3) of the AquaControl simulations. Black dashed line indicates the cluster mean of the respective network measure. Vertical lines indicate the time-mean position of the ITCZ in the respective models.

has been explained by the strong southern Hadley cell and the resulting enhanced propagation of SST anomalies. In contrast to this observation, models in cluster 2 (second row in Fig. 6.4) are characterized by a maximum (in the average link distance) or at least elevated values (in the degree) around the equator. Simultaneously, models in cluster 2 show rather symmetric network measure distributions with respect to the equator. We conclude that these observations (more symmetric network measures and elevated values around the equator) are correlated with the smaller northward shift of the ITCZ and a weaker SST contrast.

The single model clusters 3+4 directly follow this rationale. Whereas cluster 3 (Caltech model) shows almost perfect network measure symmetry and, thus, a near equator ITCZ and low SST contrast, the AM2.1 model in cluster 4 exhibits a clear maximum at the position of the southern Hadley cell indicating the far northward shifted ITCZ and the corresponding large SST contrast.

To further shed light on the correlation between the zonal network distribution on which we have based the clustering analysis, we investigate the elements of the similarity matrix $\boldsymbol{S}$. As explained in the methodology, the matrix elements s_{ij} represent the sum of the pairwise similarity between the average link distance and degree distributions and are, therefore, bounded by $[-2, 2]$. In Fig. 6.5, we show the matrix elements and the corresponding mean ITCZ position difference (marked by $+$) and SST contrast difference (marked by $\cdot$) as a scatter plot. In addition, we have colored the pairs, which have been clustered together in red (cluster 1) and blue (cluster 2).

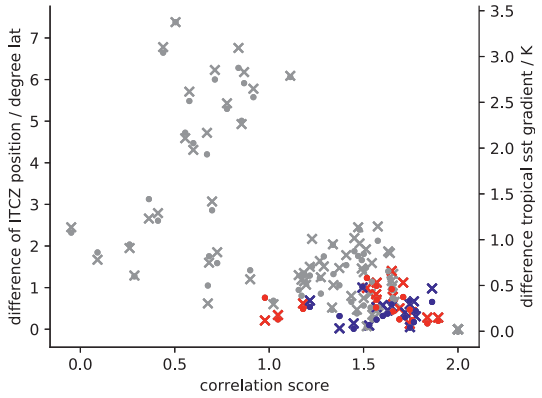

Figure 6.5.: Matrix elements (correlation scores) indicating the network measure similarity between model pairs versus the difference in time-mean ITCZ (+) and time mean SST contrast (·). Models pair within one cluster are marked red (cluster 1) and blue (cluster 2).

Figure 6.5 mainly highlights two observations. First, there is a trend of smaller SST contrast difference (and mean ITCZ position difference) for models with higher similarity. This can directly be interpreted as a confirmation of the analysis framework. We have shown that zonal mean network measures, which are based on the correlation between SST anomaly time series are closely related to the mean position of the ITCZ and the tropical SST contrast. Second, we observe that the colored elements in the scatter plot are located in the bottom right corner which is characterized by low SST contrast difference and high similarity between the model pairs. We interpret that as a confirmation of the clustering method and conclude that the analyzed clustering has been meaningful.

6.5. Network analysis - excluding extratropic-extratropic connections

To disentangle the importance of different global connectivity patterns regarding their influence on the mean ITCZ position, we stepwise exclude connectivity patterns and analyze the resulting networks with the described framework.

As a first step, we delete all cross-equatorial edges connecting the respective extratropics ($> 35°$N/S). This results in almost the same networks (no visual difference in network measures and identical clustering). We, therefore, conclude that potential cross-equatorial teleconnectivity patterns do not significantly influence the ITCZ position or are not present in the monthly SST variability of the aquaplanet scenario.

In a second step, we exclude all extratropic-extratropic connections. In this case, it is important to emphasize that excluding connections in a network is realized by removing edges but not nodes. As a result of the exclusion scheme, which we achieve by manipulating the adjacency matrix, nodes in the extratropics exhibit a strongly

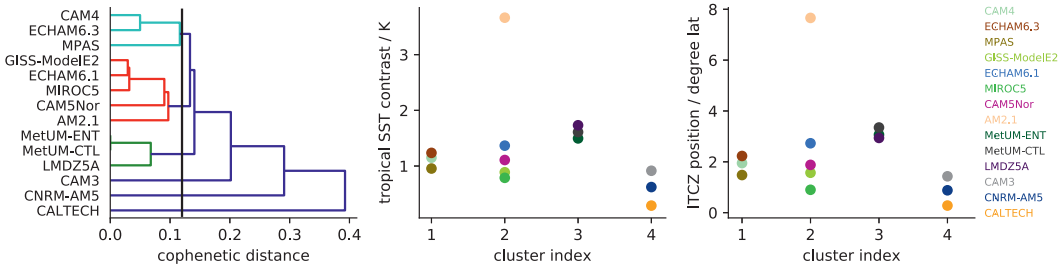

Figure 6.6.: (a) Dendrogram based on the clustering of the zonal-mean network measures utilizing global connection patterns excluding extratropic-extratropic links. Arrangement of panels as in Fig. 6.3.

decreased degree. Accordingly, an analysis following in the previously described way leads to similarity scores between the manipulated distributions which are heavily influenced by the values in the extratropics. This phenomenon is amplified by the fact that the extratropics contribute 110 latitudinal degrees to the whole latitude range whereas the tropics contribute 70 degrees latitude only. Therefore, we need to adapt the analysis framework to access the influence of tropical-extratropical connection on the mean ITCZ position.

In networks, edges contribute to the degree of two nodes. Therefore, the degree and average link distance of the nodes in the tropics depend on the edges to the extratropics. Here, we make use of this by only analyzing the degree and link distance of the nodes in the tropics of the network in which we have excluded extratropic-extratropic edges. With this strategy, we achieve both, independency from the biased values in the extratropics and a simultaneous dependency on the links between the extratropics and the tropics. Surprisingly, we can also realize this by utilizing the part in the tropics (35°S to 35°N) of the zonal mean network distributions from the previous section, as these values are not dependent on extratropic-extratropic connections.

Selecting this latitude range from the previous section as an input for the analysis framework leads to the clustering which we show in Fig. 6.6a. In this case, the 14 GCMs get clustered in four groups of similar size, which to some extent also separate regarding their tropical SST contrast and mean ITCZ position (6.6b,c).

In particular, models in cluster 1 exhibit a suppressed northward shift of the ITCZ as well as a reduced tropical SST gradient in comparison to cluster 3. In contrast, cluster 2 is characterized by widespread mean ICTZ positions and SST gradients. Finally, models in cluster 4 (which is a group of three single model clusters) all show very small SST gradient and ITCZ positions close to the equator. We especially emphasize the alignment of the values in cluster 4 which seem to contradict the observation of large cophenetic distance.

To further understand the cluster allocation and, specifically the interpretation of cluster 4 as a distinct group, we show the zonal mean network measures in the corresponding range between 35°S and 35°N in Fig. 6.7. Comparing cluster 1 and

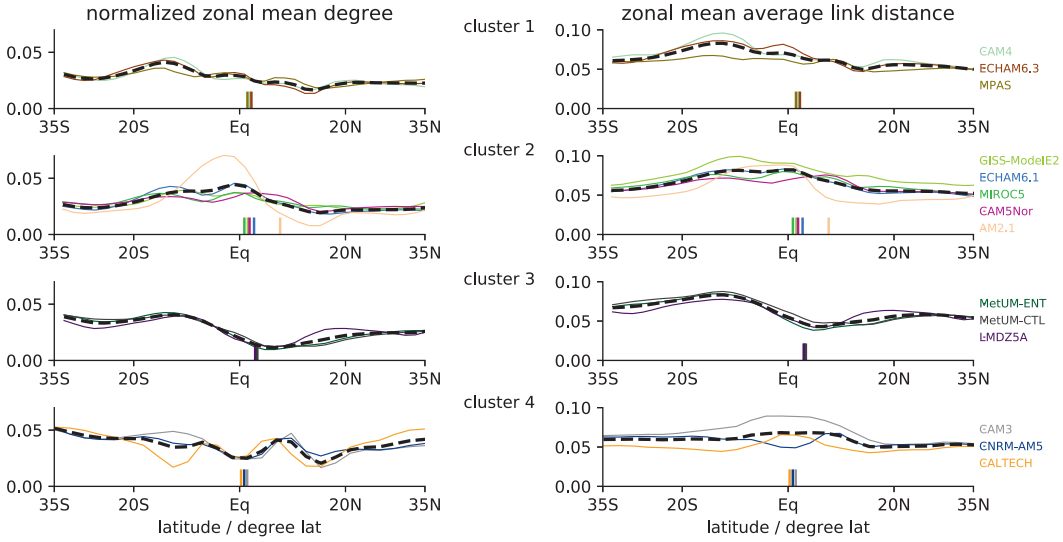

Figure 6.7.: Zonal mean average node degree (left panels) and zonal mean average link distance (right panels) for the previously identified clusters (vertically ordered, see Fig. 6.6) of the AquaControl simulations excluding extratropic-extratropic connections. Arrangement of panels as in Fig. 6.4.

cluster 3 (row 1 and 3 in Fig. 6.7) which separate well by the means of SST gradient and ITCZ position, we find that the respective zonal mean network distributions are all characterized by a maximum of the zonal mean degree at the center of the southern Hadley cell. Besides, we observe a minimum of both network measures north of the position of the mean ITCZ. The zonal mean network measure distributions differ in mainly two points: first, the described minimum of the network measures is shifted further northwards for cluster 1 models and their degree and average link distance distribution are additionally characterized by two local minima between the respective minimum north of the mean ITCZ position and the global maximum on the Southern Hemisphere. We assume, that these additional features are sourced by some atmospheric wave pattern suppressing the cross-equatorial energy transport and, thus, restrain a further northward shift of the ITCZ and a greater SST gradient.

In line with the observation of widespread SST contrasts and ITCZ positions of models in cluster 2 (see 6.6b,c) also the zonal network measures of the models of this cluster exhibit a greater spread and do not allow for coherent analysis. Similarly, cluster 4 is also characterized by dissimilar zonal network measures which lead to the large cophenetic distance between these models. But, we notice that all models exhibit fairly symmetric network measures that agree with the observation of small tropical SST contrasts and ITCZ positions close to the equator.

In summary, we have shown that the symmetric network measure distributions are coherently linked with small SST gradients and near equator ITCZ positions and vice versa in both network setups studied. In addition, we have studied several other network configurations (e.g. using tropical-tropical connections only). These analyses

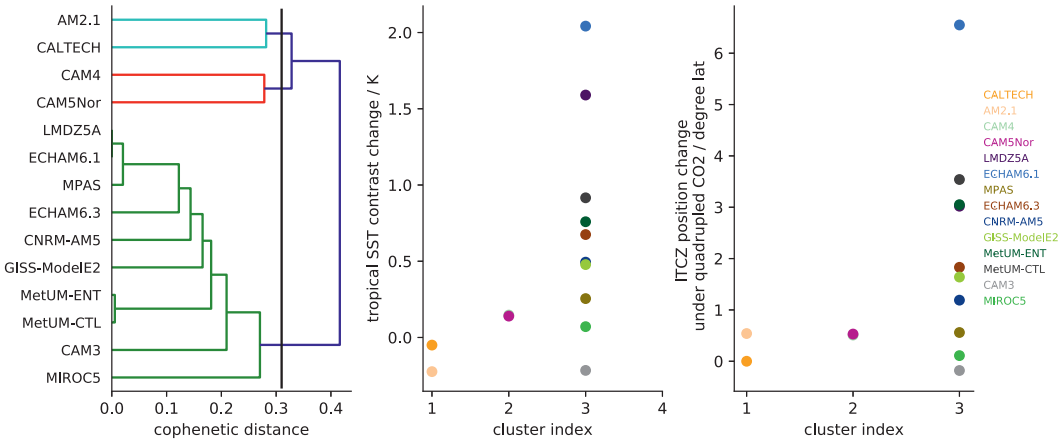

Figure 6.8.: (a) Dendrogram based on the clustering of the zonal-mean network measures utilizing the difference between the zonal mean network measures based on the AquaControl and Aqua4xCO2 model runs. Arrangement of panels as in Fig. 6.3.

have not led to useful insights and especially not to meaningful model clusters. We, therefore, conclude that considering tropical-extratropical interactions is crucial to understand the ITCZ dynamics which highlights the global character of the ITCZ and has also been shown by a recent study [142].

In the following, we proceed by analyzing networks based on the Aqua4xCO2 model runs. In these scenarios, we aim for understanding the climate change response of the ITCZ triggered by the elevated carbon dioxide levels.

6.6. Network analysis - climate change response

To test the functionality of the framework to identify systematic differences among groups of models regarding their global warming response that is induced by the quadrupled carbon dioxide concentration, we construct networks based on the Aqua4x-CO2 model runs. To evaluate the model-specific differences between the networks based on the Aqua4xCO2 model runs and the AquaControl model runs and correlate these with the respective response in the SST gradient and mean ITCZ position, we subsequently subtract the zonal mean network measure based on the AquaControl model run from the zonal mean network measure in the climate change scenario. We use this difference as input for the clustering framework.

Figure 6.8 shows the clustering of the models and the respective SST gradient and mean ITCZ difference between the AquaControl and the Aqua4xCO2 scenarios. The framework identifies three clusters where cluster 1 and cluster 2 consist of two models each. Although cluster 1 and cluster 2 separate well in terms of SST contrast change and ITCZ position difference, we do not consider this clustering as successful, as most models are grouped in cluster 3. In addition, cluster 3 features a broad variety of climate change responses. In general, the climate change response is heterogeneous

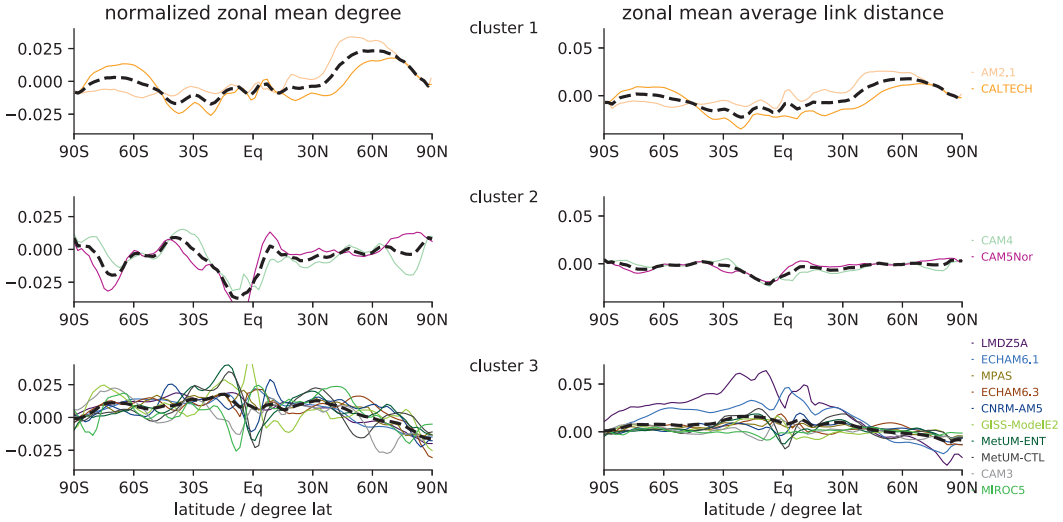

Figure 6.9.: Difference between the zonal mean average node degree (left panels) and zonal mean average link distance (right panels) of the AquaControl and Aqua4xCO2 simulations for the previously identified clusters (vertically ordered, see Fig. 6.8). Arrangement of panels as in Fig. 6.4.

among the models. We observe slight reductions of the tropical SST gradient as well as an increase of 2°K. The mean ITCZ position changes in the range from a small southward shift to a strong northward shift of nearly 7° latitude.

To shed light on the zonal mean network measure differences that result in the above-described clustering, we show the corresponding distributions in Fig. 6.9. The first two clusters both exhibit a muted response to the drastic increase of carbon dioxide. Although the distributions in clusters 1 and 2 have only a few features in common, we still observe a coherent increase or constant values of the network measures in the Northern Hemisphere. In contrast, most models in cluster 3 show reduced degree and link distance on the Northern Hemisphere which we link with suppressed atmospheric dynamics in this latitude range. As the zonal mean network measures otherwise substantially differ and show almost no agreement, we conclude that the framework is not able to meaningfully separate models with distinct climate change response and corresponding similar change in ITCZ dynamics.

6.7. Robustness of the results

We are aware that although our analysis setup has been specifically designed for studying zonal network measure distributions, there exist alternative strategies to tackle the stated research question. To validate the significance of our clustering analysis, we have conducted two tests. On one hand, we have split the 30 years of the in-equilibrium time series of the 14 AquaControl model runs into two parts of 15

years length. An application of the analysis framework leads to two clusterings of the models which only slightly differ from the result we have presented above.

In addition, we have randomly selected and shuffled 20 years of the same time series for each model and have performed 20 iterations of this procedure. For each model, we have compared the resulting mean of the zonal mean network measures with the original zonal mean network measure distribution. We notice that the main features of the zonal mean network distributions agree.

Considering all the above we conclude that our results are sufficiently stable.

We have attached the corresponding figures for this section in the appendix to this chapter (Appendix C).

6.8. Discussion and summary

In this chapter, we have applied the framework of functional climate networks to investigate the signature of ITCZ dynamics in 14 global circulation models. In particular, we have clustered the models by their zonal mean network measure distributions. Utilizing the AquaControl model runs and hierarchical clustering we have identified meaningful clusters considering the global network and a manipulated network with deleted extratropic-extratropic connections. The clusters in both above-mentioned cases separate regarding their mean ITCZ position and tropical SST contrast. Considering this result, we conclude that we have, in contrast to previous approaches [131–134], linked time mean ITCZ dynamics with monthly SST variability patterns. This result illustrates the power of network-based approaches to extract additional information from climate data which complement traditional analyses using self-organizing maps and empirical orthogonal functions. As complex networks have not yet been applied to study, validate, and cluster output from global circulation models, we interpret this study as a step towards data-driven techniques in this direction.

As we have limited our analysis to the study of zonal mean network measures, we have neglected information hidden in the meridional links in the network. A more detailed investigation of the two-dimensional network pattern can possibly reveal the physical mechanisms leading to the described differences between the models and is outlined as future research.

Part III.

Complex networks in Computational Social Science

In the two chapters of this third part of the thesis, the dynamically evolving popularity of content and the individual behavior on social media platforms are examined. By analyzing the topology of content networks which are based on the co-occurrence of hashtags in chapter 7, it is shown how distinct trends are connected via related buzzwords. To study the temporal evolution of these content networks, a comprehensive matching scheme is developed. The usability of this approach is illustrated by accessing the bursty popularity dynamics in a fashion blog. In chapter 8, the findings related to the research question "How the social media platform Twitter has changed in the last decade and how Twitter has potentially changed us?" are explored. This is done by tracking the tweeting behavior of $600,000$ users over the time of 7.5 years and evaluating the individual behavior at different stages of the Twitter career.

Chapter 7.

Capturing the popularity of hashtag communities

7.1. Introduction

Online platforms that provide interactive environments play a prominent role in modern societies [38, 43, 143, 144]. Due to their omnipresence, they have gained attention by research and economy [4]. Especially, the analysis of shared content has become crucial to understand information diffusion in the online world. This work has led to new insights about the formation of trends, ways of interaction, and opinion formation [143, 145, 146].

With the rise of social media, hashtags have been invented to easily find related posts and connect them to different topics and trends. Hashtags are built up by the combination of the #-symbol with a number of letters, which can be anything ranging from abbreviations of slogans (as in e.g. #ootd standing for *outfit of the day*) to whole sentences (as in #blacklivesmatter). With the emergence of topics, hashtags are simultaneously being created, become popular in this context, and can often be uniquely attributed. Therefore nowadays, hashtags are an established way to link descriptive buzzwords to posts and set the shared content in a context of some ongoing discussion, trend, or topic on many different platforms [146, 147].

Here, we propose a methodological framework that is able to capture this dynamically changing stream of topics and trends. For this purpose, we set up a network where nodes of the networks represent hashtags which we cluster into groups. We approximate these groups as distinct topics.

One way to efficiently group nodes in networks is community detection, a type of unsupervised clustering in networks. Such communities can exhibit different characteristics and can be obtained by the application of various detection algorithms [61, 68, 148, 149]. In the last two decades, the improvement of community detection in static networks has gained attention [64]. Naturally, also tracking the evolution of communities in temporal networks is a subsequent research question which has been addressed by different authors [39, 150, 151]. Here, we tie in with other studies which promote approaches using comprehensive matching schemes to connect communities from subsequent static network snapshots [60, 152, 153]. In contrast to other approaches, our algorithm is independent of the static community detection algorithm and, therefore, broadly applicable.

In this chapter, we initially propose a way to construct networks from hashtag co-occurrences. By analyzing the static networks from weekly hashtag co-occurrences in posts on the fashion blog `lookbook.nu`, we find that hashtags are hierarchically ordered and communities exhibit mainly two characteristic topological patterns. To enable the analysis of the temporal change of the networks' community structure, we introduce a matching scheme to stack the subsequent weekly network slices. Our approach is based on a combination of different algorithms and incorporates higher orders of memory [154]. Finally, we illustrate the functionality of the framework by analyzing the dynamic popularity landscape of co-occurring hashtags on `lookbook.nu`.

The results presented and the figures shown in this chapter are based on publication P_6 (*Lorenz-Spreen, P., Wolf, F., Braun, J., Ghoshal, G., Djurdjevac Conrad, N., Hövel, P. (2018): Tracking online topics over time: understanding dynamic hashtag communities, Computational Social Networks 5.1*). We thank Springer for the kind permission to reuse/adapt the content and figures under the creative common license `http://creativecommons.org/licenses/by/4.0/`.

7.2. Data and network construction

Social platforms have often the appearance of a network by nature as people interact via (usually bidirectional) friendship connections or (usually unidirectional) follower-following connections [35]. This results in undirected and directed network structures, respectively. In contrast to the rather obvious user interaction network, we here suggest to base networks on the co-occurrence of hashtags that users utilize to tag their posts. This procedure has previously been employed to investigate social tagging systems [155]. To construct a network representing the connection between different topics, we interpret hashtags as nodes in a network and connect these nodes if a pair of hashtags is simultaneously used in some user's post. The basic idea is illustrated in Fig. 7.1, where we show screenshots of two posts on the fashion platform `lookbook.nu` and the resulting network. Due to the proposed edge definition, we obtain a complete subgraph of size n with every post that contains n different hashtags. By aggregating this stream of small networks that are based on the different posts into a single network, we obtain a weighted topic-relation network. As we are specifically interested in the evolution of topics we aggregate posts within distinct time windows into different networks. We finally interpret the temporal stream of static networks referring to the post activity within the distinct time windows as layers of a multiplex network which enables us to temporally study relations of topics and their relative importance.

As already mentioned, we utilize a dataset from `lookbook.nu` which has been obtained in May 2017 by an HTML scraper. Lookbook is a fashion platform where users can post so-called *looks* (fashionable pictures) which are often tagged using more or less specific hashtags and connect via unidirectionally *following* each other. To show appreciation of a post users can *hype* posts. Screenshots of typical posts are partially shown in Fig. 7.1. Its architecture is similar to other popular social

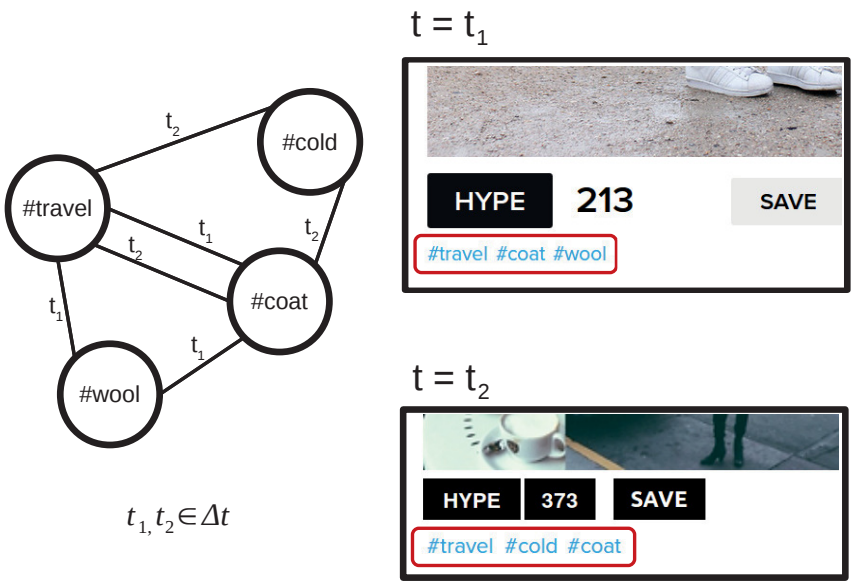

Figure 7.1.: Construction of hashtag co-occurence networks. Hashtags which appear under posts (screenshots on the right) get connected in a temporal network (shown on the left).

media platforms such as *Instgram* which makes Lookbook an interesting case for studying social networks. Crawling along the follower-following connections, the content of 22,748 users and 1,158,340 associated posts containing 81,409 individual hashtags have been acquired. The above-described procedure leads to 1,358,241 weighted edges (hashtags appear together under multiple different posts) connecting the different hashtags. For our analysis, we choose an aggregation window of $\Delta t = 1$ week to neglect intra-weekly time cycles and, therefore, obtain 52 static network snapshots for the year 2015. As standard network measures for these networks (mean degree, network diameter, average shortest path length, global clustering coefficient) remain comparably constant for all 52 snapshots, we assume this choice to be of appropriate size.

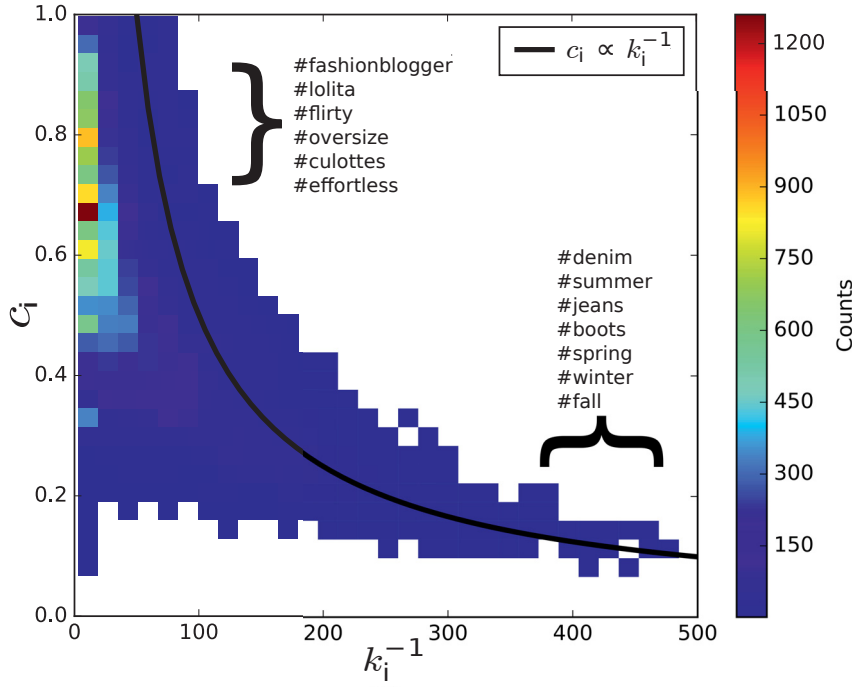

Figure 7.2.: Color-coded two-dimensional distribution of degree and local clustering coefficient. The most frequently used hashtags in the respective marked interval are listed.

7.3. Hierarchical order of hashtags and different community types

To initially examine the structure of the network, we investigate the degree and the clustering coefficient of the networks' nodes which each represent a distinct hashtag. Here, we measure the degree and clustering coefficient across all different snapshots and calculate the mean for each hashtag.

For that, we utilize the unweighted clustering coefficient but the weighted node degree

$$k_i = \sum_{i=1}^{N} w_{ij}. \tag{7.1}$$

In Fig. 7.2 we show a density plot where the x-axis represents the degree whereas the y-axis indicates the local clustering coefficient. We can observe a $c_i \propto k_i^{-1}$ dependency which is the relationship that a simple model for hierarchical networks predicts [156]. This finding can be supported by an exemplary study of the different hashtags which get each associated with distinct bins in Fig. 7.2. Whereas hashtags like #summer or #denim with an elevated degree and low local clustering coefficient are top-level hashtags of different topics, low degree and increased clustering associated hashtags

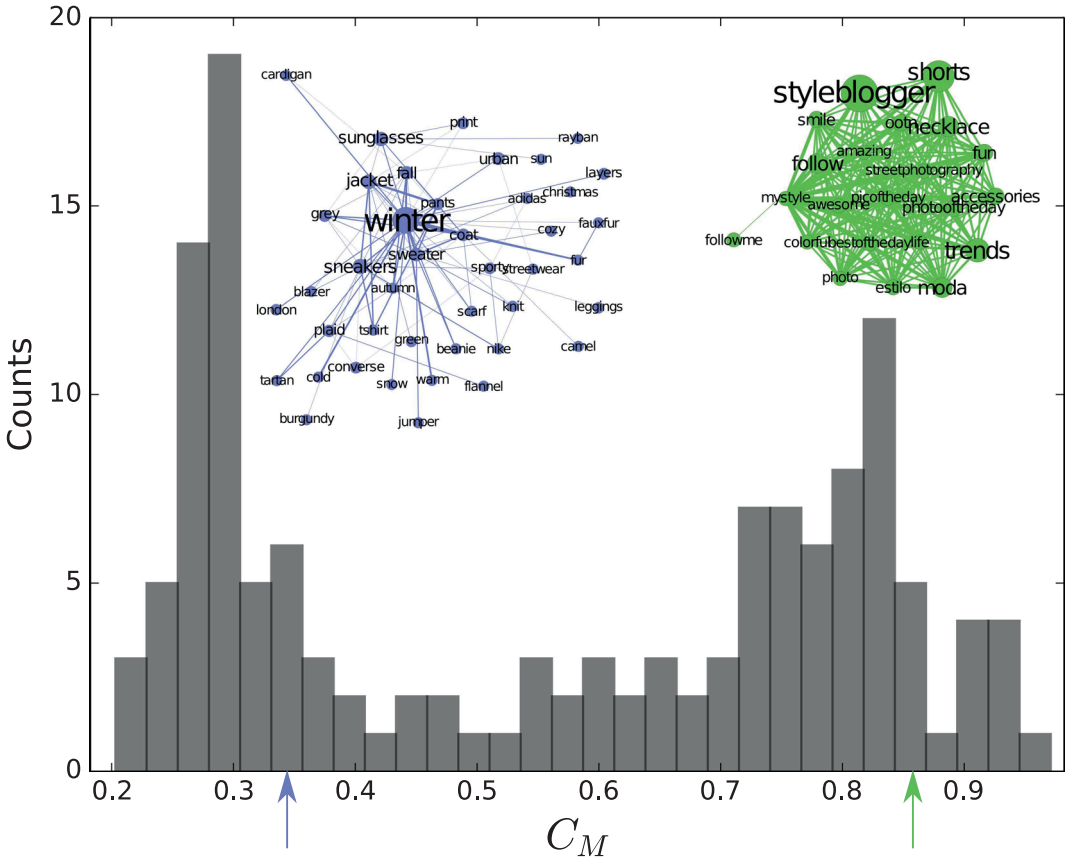

Figure 7.3.: Histogram of the module (community) mean clustering coefficient. We show two respective examples (with marked mean clustering coefficient) illustrating the features of the bimodal distribution.

such as #oversize, #culottes, or #effortless either describe more specific styles or are nonsense hashtags which can be used in many different contexts. We, therefore, do not only find topological indications but also semantic hints for a hierarchical order of hashtag usage.

To identify broader topics and groups of hashtags, we have applied different community detection algorithms for each of the static weekly-binned networks. As different community detection methods have led to topologically and visually similar results (not shown), we utilize modularity maximization (Eq. 2.14) based spectral partitioning. The relatively high modularity of all networks ($Q_t > 0.5$) enables us to interpret the networks as large, rather well-separated, selections of different topics which are each characterized by a number of more or less specific hashtags [63, 64].

For further classification of the modularity maximization based hashtag communities we calculate the mean clustering coefficient of all hashtags in each community for each of the 52 networks and show the corresponding histogram in Fig. 7.3.

The bimodal distribution of the module-mean local clustering coefficient indicates

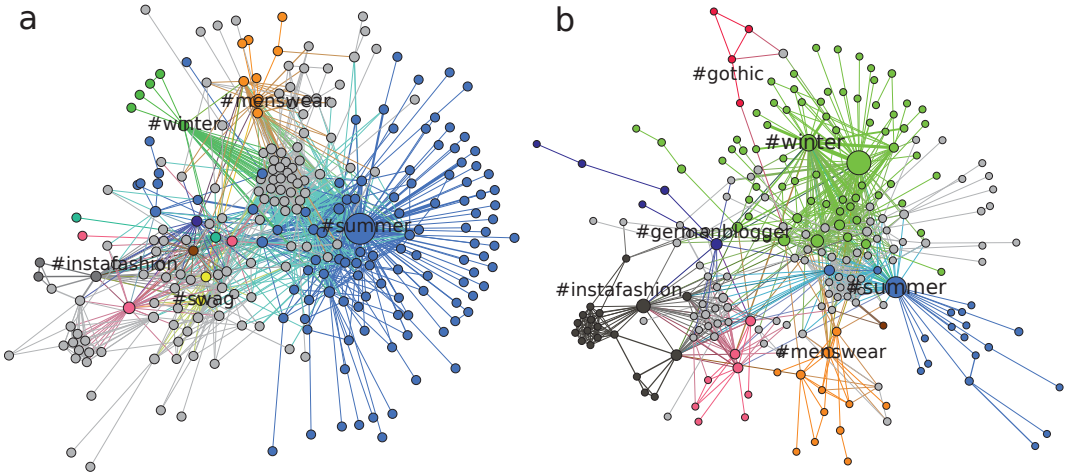

Figure 7.4.: Two snapshots of the temporal co-occurence networks (a: July, b: December). Nodes are colored according to their allocation into communities (colors) and transition regions (grey).

the existence of two main classes of communities. On the one hand, we find star-like communities with a hub and many unconnected nodes in the periphery, and on the other hand, we identify nodes with a rather flat degree distribution and dense interconnectivity. A semantic analysis of the different hashtag communities reveals that low clustering coefficient (star-like) communities like the purple community in Fig. 7.3 consist of descriptive hashtags, whereas the usage of rather nonsense buzzword hashtags leads to communities with elevated clustering coefficient.

7.4. Tracking temporal community evolution

7.4.1. Distinguishing temporal communities

Networks from different time windows are topologically similar, but differ in present nodes (used hashtags in the specific period) and, thus, also their community structure. In Fig. 7.4 we show two different snippets of networks from August and December. The different colors of the nodes indicate the community association. To filter out the rather unspecific hashtags which can be used in multiple contexts and, therefore, do not belong to the core of some community, we utilize a suited random walker based community detection algorithm [66, 67] to identify nodes belonging to transition zones (colored in grey). The details of this approach are explained in detail in publication P_6.

Comparing the communities in Fig. 7.4a and Fig. 7.4b which are based on hashtag co-occurrences in July and December, respectively, we observe a shift from a major community associated with the hub #summer to a dominant community with #winter in the center. As `lookbook.nu` is a global platform (which is dominated by users from the Northern Hemisphere), we can still identify the opposite community as a small

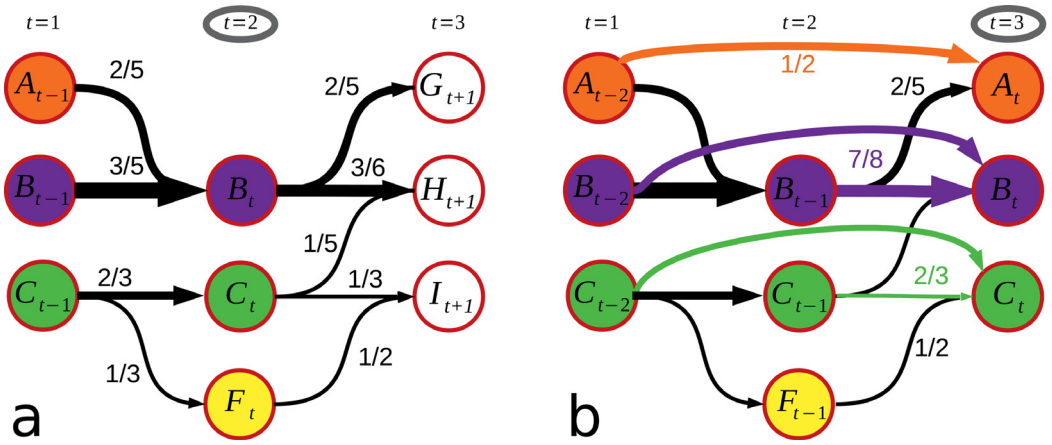

Figure 7.5.: (a,b) Illustration of two exemplary steps of the temporally extended Jaccard-Index based matching scheme. Colors indicate matched communities, fractions indicate overlap. Colored fractions indicate the memory weights leading to the according matching.

group of nodes in the other network. In addition to that, we find both, communities that are present in July and December (e.g. the community around #instafashion) but also appearing and disappearing communities (e.g. related to #menswear or #swag).

To track the community evolution, we consecutively stack the static network layers of the weekly hashtag co-occurrences and perform a multilayer comparison algorithm to match communities from different time windows. We introduce this strategy as a more general tool in the following.

7.4.2. Memory weights

To match network parcellations of different networks we suggest considering the Jaccard-Index

$$I\left(A_{t-1}, B_t\right) = \frac{|A_{t-1} \cap B_t|}{|A_{t-1} \cup B_t|}, \tag{7.2}$$

with labels A_{t-1} and B_t of two different communites from subsequent time steps.

In this definition, the Jaccard Index is a measure of the overlap of two communities. Due to the nature of community detection algorithms, network parcellations are often unstable and tend to vary for different parameter settings or small topological changes of the network. In a temporal evolution of network communities, we expect communities to grow and shrink but also to merge, split, and reunite. To account for the instabilities that come along with static community detection, we propose the *memory weights* as a temporal extension of the Jaccard-Index

$$W\left(\{A_{t-n},...,A_{t-1}\},B_t\right)=\sum_{t'=1}^{n}\frac{1}{t'}\frac{|A_{t-t'}\cap B_t|}{|A_{t-t'}\cup B_t|} \tag{7·3}$$

Here, we measure the overlap of a community of the present time-step with all temporal instances of one community A of the previous time steps. In the version stated above, we introduce the simple memory kernel $\frac{1}{t'}$ which can, in principle, be replaced by different choices. Exemplarily, utilizing an exponential memory kernel

$$W\left(\{A_{t-n},...,A_{t-1}\},B_t\right)=\sum_{t'=1}^{n}e^{-t'\cdot r}\frac{|A_{t-t'}\cap B_t|}{|A_{t-t'}\cup B_t|} \tag{7·4}$$

allows to estimate the memory based on the mean relative overlap of all nodes (or hashtags) of adjacent network parcellations.

To illustrate the memory weights approach, we execute the relabeling of communities with the help of examplary communities from an imaginary network parcellation in Fig. 7.5. In the second time step of a relabelling process, we first need to apply a matching technique. Due to the fact that we can only compare the communities at $t = 2$ with one previous set of communities, we measure the overlap and rename the communities according to the maximal overlap. As relabelling needs to be unique, left communities (here community F_t in yellow) appear as new communities and do not get relabelled. At $t = 3$ we apply the memory weights to match the newly found communities with the already identified communities. The proposed method (with the memory kernel $\frac{1}{t'}$) allows for rediscovering community A as $W\left(\{A_{t-1},A_t\},G_{t+1}\right)=0+\frac{1}{2}$ is larger than $W\left(\{B_{t-1},B_t\},G_{t+1}\right)=\frac{2}{5}+0$. In addition, we correctly identify community I_{t+1} with community C as the overlap between $t = 1$ and $t = 2$ is considered and hence $W\left(\{C_{t-1},C_t\},I_{t+1}\right)=\frac{1}{3}+\frac{2}{6}>W\left(\{F_{t-1},F_t\},I_{t+1}\right)=\frac{1}{2}+0$.

To correctly realize the label allocation by maximizing the overlap between the actual and the previous community parcellations, we utilize the Hungarian algorithm [157]. This algorithms solves the bipartite matching problem in polynomial time.

Finally, note that the introduction of a lower limit of a matching score should be applied to avoid unintended allocation of community labels. In addition and most importantly, the proposed matching scheme is independent of the utilized static community detection method and, therefore, is broadly applicable.

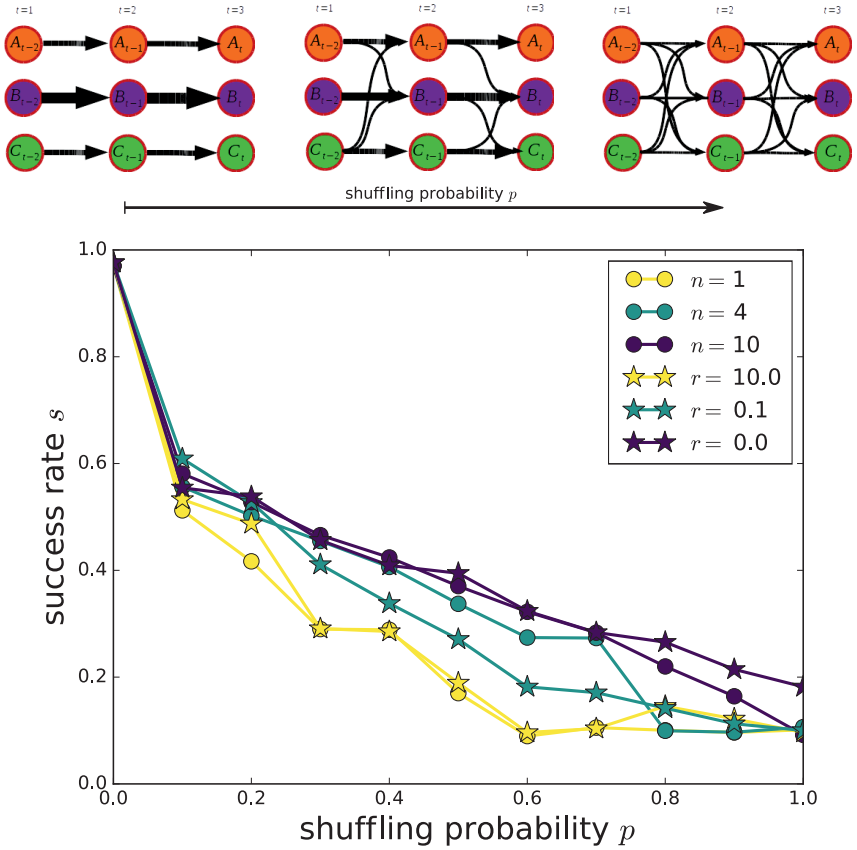

Figure 7.6.: Stability test of the matching algorithm. Circles indicate n-th order matching, stars represent an exponential kernel with exponent r.

7.5. Stability test of the proposed method

Next to the motivation of introducing memory to mimic fading long-term correlation of similar hashtag use, the memory weights also inherit the advantage of stabilizing possible uncertainties of static community detection. To demonstrate this feature of the proposed combination of algorithms, we construct a test case where we pertubate a defined test community structure with temporally uncorrelated noise. This setup imitates possible misallocations due to the static community detection.

As illustrated in Fig. 7.6, we shuffle community members at random with shuffling probability p and construct randomized copies. Following this procedure, we produce a surrogate multilayer community time series of length 10. To test both introduced memory kernels (Eq. 7.3 and Eq. 7.4), we subsequently apply the above-introduced scheme in three parameter settings ($n = 1$ and $r = 10$: no memory, $n = 4$ and $r = 0.1$: intermediate/realistic memory, $n = 10$ and $r = 0$: infinite memory) and compare the resulting community labelling with the benchmark parcellation. Finally, we compute the success rate s (correct labeling vs. all other incorrect labeling outcomes) for

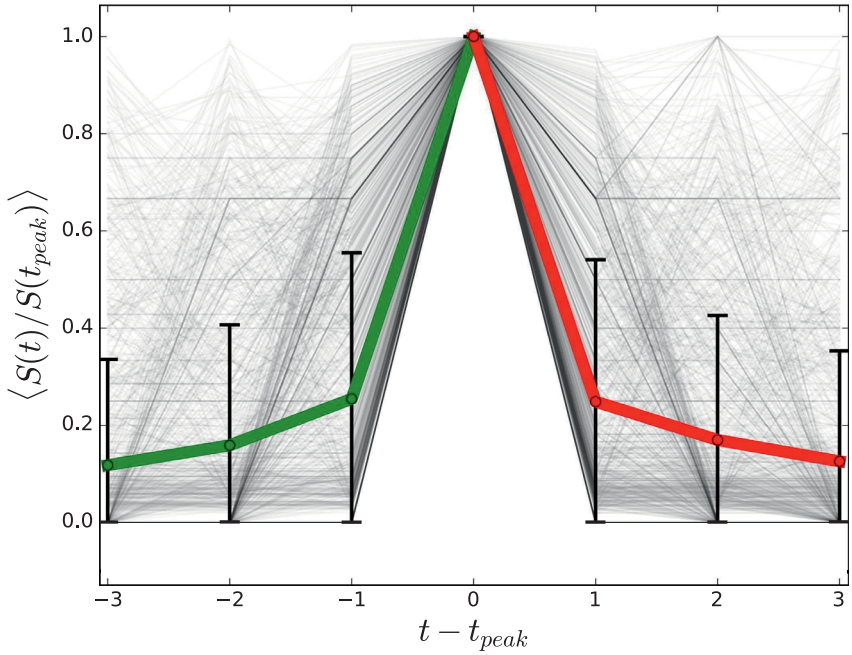

Figure 7.7.: Relative size before and after the size maximum. The mean trajectory is marked in green (gains) and red (losses). Bars indicate the variance.

different shuffling probabilities p and show the result in Fig. 7.6.

We observe that the success rate decreases, as expected, with increasing shuffling probability but can be improved by increasing the memory of the matching scheme. Whereas the memory weights in the form as in Eq. 7.3 with $n = 1$ correspond to the application of the Jaccard-Index and perform worst (measuring the pairwise overlap between subsequent time steps only), increasing the memory to $n = 4$ already clearly improves the success rate. We find similar results for an exponential kernel. Furthermore, for low shuffling probabilities, there is only a small difference between intermediate and infinite memory indicating respective advantages for real-world applications where we expect noise in the low shuffling probability regime.

7.6. Temporal hashtag communities

Having the methodology set up and the algorithm tested, we can finally analyze temporal communities identified by employing modularity maximization (Eq. 2.14) within the hashtag co-occurrence network of `lookbook.nu`.

Based on the mean hashtag overlap between subsequent complete networks $\frac{H_t \cap H_{t-1}}{H_t \cup H_{t-1}} \approx 0.9$ we choose an exponential kernel with $r = 0.1$ for our analysis.

An application of the memory-based community relabelling enables us to track the growing and shrinking of communities. In the following, we interpret the size $S(t)$ of a community (the number of members) as a proxy for the popularity of a topic

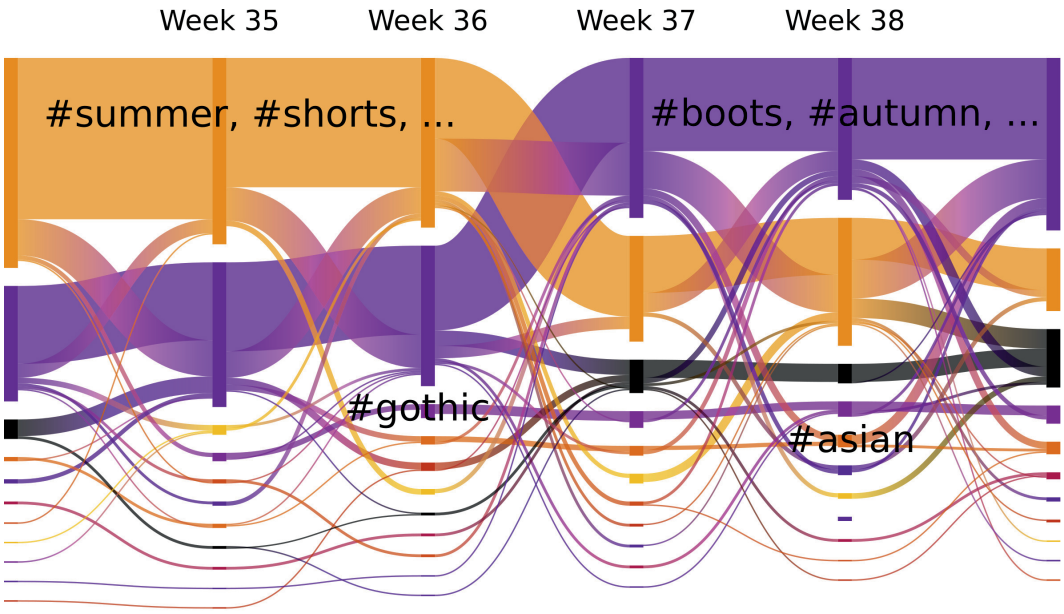

Figure 7.8.: Alluvial diagram of the dynamic hashtag communities between week 34 and week 39 of year 2015. Vertical bars repesent community size, horizontal bars shared members of subsequent communities.

that is described by the participating hashtags. To analyze how topics alter in their popularity, we investigate the mean trajectory of each community before and after its relative peak $S(t_{\text{peak}})$.

Figure 7.7 shows the ensemble of the resulting relative increases and decreases of community sizes with the mean increase marked in green and the mean decrease highlighted in red. Figure 7.7 illustrates the multiple dynamic features of community emergence in the `lookbook.nu` data set. All communities exhibit dynamic size shifts and a broad variety of different trajectories around the respective peak. The increasing variance close to the maximum at $t = t_{\text{peak}}$ indicates the various different possible trajectories around the peak ranging from longer-lasting plateaus to short bursty popularity periods. In addition, we observe a striking similarity between size (or popularity) gain and loss in Fig. 7.7 implying that not only popularity gain but also popularity loss is distributed broadly. The burstiness is illustrated as often more than 70% of the maximum size is gained (lost) in the time step before (after) the peak. In the meantime, this behavior has been shown to be characteristic of online popularity and has become even more bursty in recent years among various data sets which is suspected to be highly linked to an acceleration of online discourses [147].

To visualize the shifted developments and bursty size gains of the weekly communities, a snapshot of the corresponding alluvial diagram [40] is depicted in Fig. 7.8.

In the alluvial diagram Fig. 7.8, the number of communities is reflected by the size of the vertical bars below the annotation of the time step whereas the communities

themselves are sorted by their size. Horizontal bars between communities from different time steps indicate shared members. The coherent coloring of the community has been realized using the memory weights in the described parameter setting.

Figure 7.8 depicts the bursty popularity shifts of the different communities. We observe a season-related change of the biggest community related to summer which gets replaced as the biggest community by the community related to autumn between week 36 and week 37 (first weeks of September 2015). Simultaneously, smaller communities merge, shift their rank, or disappear while some small communities related to specific groups (e.g. #gothic, #asian) persist and also form stable communities. An interactive tool has been programmed by Philipp Lorenz-Spreen and can be accessed at `www.tu-berlin.de?lorenz`.

7.7. Summary

Temporal community detection is prone to many different challenges, such as instability with respect to small topological changes. In this chapter, we have not only proposed a method to conduct temporal community detection but have also analyzed the evolution of trends in a fashion blog.

Based on the co-occurrence of hashtags, we have constructed weighted networks out of posts on a fashion blog. By accumulating co-occurrences on a weekly basis, we have constructed 52 temporally ordered snapshot networks. The node degree and clustering coefficient exhibit a $c_i \propto k_i^{-1}$ dependency indicating a hierarchical order of hashtags. This finding has been supported by analyzing the content of the differently characterized hashtags. The community-mean local clustering coefficient is distributed bimodally highlighting two preferred community organizations, a star-like structure, and a hub-less highly connected node grouping.

Motivated by significant differences between networks from different time steps, we have investigated the temporal evolution of the communities. Independent from the applied static community detection algorithm, we have proposed a temporally extended Jaccard-Index, the so-called memory weights, in combination with the Hungarian algorithm to solve the bipartite matching of subsequent sets of communities. Thereby, we have introduced a higher-order memory to the community allocation and have shown that memory effects can stabilize and increase the potential of temporal community analysis. Thus, they help to overcome the inherent instability of static community detection. As the proposed approach is independent from the employed static community detection algorithm it is widely applicable. As the static instances do not even neccessarily have to originate from network parcellations, we can also analyze outcomes from other temporal clustering analyses with the memory weights.

By applying the proposed method, we have identified bursty size shifts of hashtag communities. Interpreting the size of a community as a proxy for the popularity of the respective topic has enabled us to study content dynamics around the popularity peak. We have shown that popularity gain and loss exhibit striking similarity and are distributed broadly among the communities. Finally, we have introduced an alluvial

diagram illustrating the dynamic evolution of hashtag communities and highlighting the stability and instability of differently characterized topics.

Chapter 8.

Cohorts of Twitter users: increasing engagement and shrinking content horizons

8.1. Introduction

Unlike in the previous chapter where we have aimed for understanding the content dynamics on social media, we, here, focus on the individual behavior of users on one of the most widely used social media platforms worldwide.

Engaging with social media has become a part of daily life in the last decade. In many European countries, more than 50% of the population is actively using social media [158]. Nowadays, 90% of the population in Europe and North America spend time online every day. In addition, more than 50% of the world's population has used online services [159].

Since internet access for households is a fairly new phenomenon, we now observe the first generation of young adults who belong to the group of "digital natives" [160, 161]. For these people, globally connected devices are the standard and have been around for their whole life. Not only for this generation but also for large fractions of global societies, sharing content online via applications on their digital devices is one important way of interacting with other people [162–165]. Therefore, technical innovations have led to a new era of communication where people connect on social media platforms worldwide [1, 166]. It is obvious that the new channels of interaction have changed, are changing, and will change how we, as humans, communicate [147, 167]. Additionally, social media platforms transform over time as well as they lose or gain popularity [168, 169].

Previous work has confirmed that online platforms influence societal opinion formation [44, 144, 159, 170]. Next to studying the topology of the networks of social media [53, 90, 146], investigating the rise and fall of topics and trends [143, 171], or discussing problematic developments and suggesting possible interventions [172, 173], current research tries to fill the gap of long-term studies on social media [174, 175].

Since the most popular social media platforms have only been around for some years to a good decade, long term studies have just recently been conducted. For studying social media interactions, the microblogging service *Twitter* has been a

frequent choice, as it is used worldwide since 2006 and reduced public data sets are available through the Twitter API.

A recent study [147] has provided first empirical evidence for a collective acceleration of public discourses and confirmed a shortening of the collective attention span not only in a Twitter data set but also reviewing movie ticket sales, google books and google trends, Wikipedia, and Reddit. In our work, we go one step beyond this study and elaborate on the individual behavior of users on the social media platform Twitter. Thereby, we generalize this finding to a shortening of the individual content horizon.

We contribute to the ongoing research by leveraging a Twitter data set that includes individual identifiers to track different users over time covering the eight years between 2012 and 2019, a, relative to Twitter's existence, long period. We show that user activity has changed on an individual level, becoming increasingly frequent, socially reactive, and dense in content over the years. We also reveal that users having joined Twitter more recently have a larger probability to become highly active. Furthermore, we demonstrate that different tweeting behavior induces differing positions in the retweet network and we find empirical evidence for a shortening individual content horizon which strongly depends on the activity of a singular user.

To make our conclusions understandable for readers without social media (or Twitter) experience, we first introduce the wording which we use in this chapter. Subsequently, we introduce the data set and review our methodology. We start the analysis by defining user cohorts that have entered Twitter at different points in time and by studying their possibly distinct tweeting pattern. Next, we determine different user types with dissimilar activity and examine their composition and content sharing behavior over time. Thereafter, we construct an interaction network based on the retweets in our data set and study the centrality of the different user types in the interaction network. We, finally, shed light on the temporal topical correlation of tweet content for the different tweeting interactions before summarizing our findings.

The results presented and the figures shown in this chapter are based on publication P_7 (*Wolf, F., Lorenz-Spreen, P., Lehmann, S. (in preparation): Generations of Twitter users reveal increasing engagement and shrinking content horizons*).

8.1.1. Defining the wording

Modern social media has led to new entities and, therefore, new terms with specific meanings. In this chapter, we will often use such terms which are not common for everybody who is unfamiliar with the social media platform Twitter. We, therefore, first introduce the wording we use in the following.

- *Tweet:* Post on the social media platform Twitter that contains a maximum of 280 characters. This is normally a short text which can be accompanied by an URL, a hashtag, or a picture.
- *Follower:* On Twitter, people unidirectionally connect via "following". If user A follows another user B, A will be shown the content and activity of user B.

- *Retweet:* Post which has been originally formulated by user A and if rewteeted by user B shown to their own followers.
- *Active retweet:* A user retweets a tweet of another user.
- *Passive retweet:* A user's tweet is retweeted by another user.
- *Inter-event time:* Time between two recorded tweets in our data set.
- *Starting day:* Date of the first appearance of a user in our data set.

8.2. Data and methods

8.2.1. Data set

For our study, we utilize data from the social media platform Twitter. Implemented in 2006, the popularity of Twitter has risen over the last decade with currently more than 300 million monthly active users (`https://de.wikipedia.org/wiki/Twitter`, Oct 21, 2020). On the platform, users can connect with each other via the following-function thereby creating a directed social network. Users can interactively share content such as texts, pictures, or videos. In addition, users can like each other's posts to show appreciation of the tweeted content. As a follower of a specific user, Twitter presents the tweeted content of this user to the follower. To spread the content of other users among the own followers, users can retweet content of each other. Due to these functionalities, the directed Twitter network serves as a substrate for dynamic spreading of content along its edges, where nowadays content is shared that varies across all kinds of topics.

For quantifying developments of long-term behavior on a broader scale within social media, it is crucial to access data sets that fulfill certain conditions. Firstly, the data sets need to cover a sufficiently long period. Secondly, the data must contain information to identify a large sample of users to be able to distinguish individual behavioral changes from collective changes in user behavior. Thirdly, the data sets have to be sampled at an adequate sampling rate so that single users can be tracked over the study period.

To conduct the study, we had access to a sample from the gardenhose data set [167] which covers the period from January 2012 until June 2019. Between January 1st, 2012, and December 6th, 2016, the data set samples 10% of the Twitter traffic dropping to 1% for the rest of the study period. The data set provides full information of the tweet such as user ID, timestamp, tweet text, and the ID of the original tweeter (in case of a retweet). For our study, we selected $N = 600,000$ users who have been active in March, April, and May 2019 and tracked their behavior over the full study period. To cope with the abrupt change of sampling rate, we resampled the data from the period with the higher sampling rate to the lower sampling rate via a uniform random sampling scheme, which imitates Twitter's random sampling routine (see illustration top of Fig. 8.1). In this chapter, we show results that are based on

the sample from April containing 200,000 users. In the appendix (chapter D), we show results from the other samples to prove the robustness of our analysis and the insensitivity of the results regarding the downsampling procedure.

8.2.2. Temporal tweet correlation estimation

In the analysis of the Twitter data set, we compute the correlation between tweet time series for different time lags. To estimate the extent to which the content of the tweet time series is correlated, we first collect the weekly tweets for each user including the used hashtags. We, here, use weekly bins, as a finer binning would lead to a huge number of empty bins (or zeros in the time series) and a larger bin size would further reduce the temporal resolution. Since the users' behavior is broadly distributed from very inactive (no recorded tweeting activity for many weeks) to very active (multiple tweets or retweets every week) the tweet time series differ fundamentally. As we are specifically interested in the topical similarity of the weekly tweet time series, we consider the Jaccard-Index

$$I\left(A_t, B_{t'}\right) = \frac{|A_t \cap B_{t'}|}{|A_t \cup B_{t'}|}. \tag{8.1}$$

Here, A and B represent sets of hashtags used in different weeks indicated by t and t'.

The utilization of the Jaccard-Index enables us to compare the similarity of the categorical data encoded in the hashtags independent of the size of the set.

In our analysis, we compute the (auto-)correlation of content for the individual weekly time series as well as the (lagged) correlation between the passive retweets and the tweets. This is achieved by computing

$$U_{i,j}(\tau) = \frac{1}{N_{\text{active weeks}}} \sum_{t=0}^{T-\tau} I(h_i(t), h_j(t+\tau)). \tag{8.2}$$

Here, $h_i(t)$ represents the hashtags used by user i in week t, T stands for the period which has passed since the user has joined Twitter and $T - \tau = N_{\text{active weeks}}$.

We, in particular, do not want to bias our results towards more or less active users. Therefore, we comute the (auto-)correlation U ignoring pairs of weeks where at least one week did not contain any activity and, thus, no used hashtags.

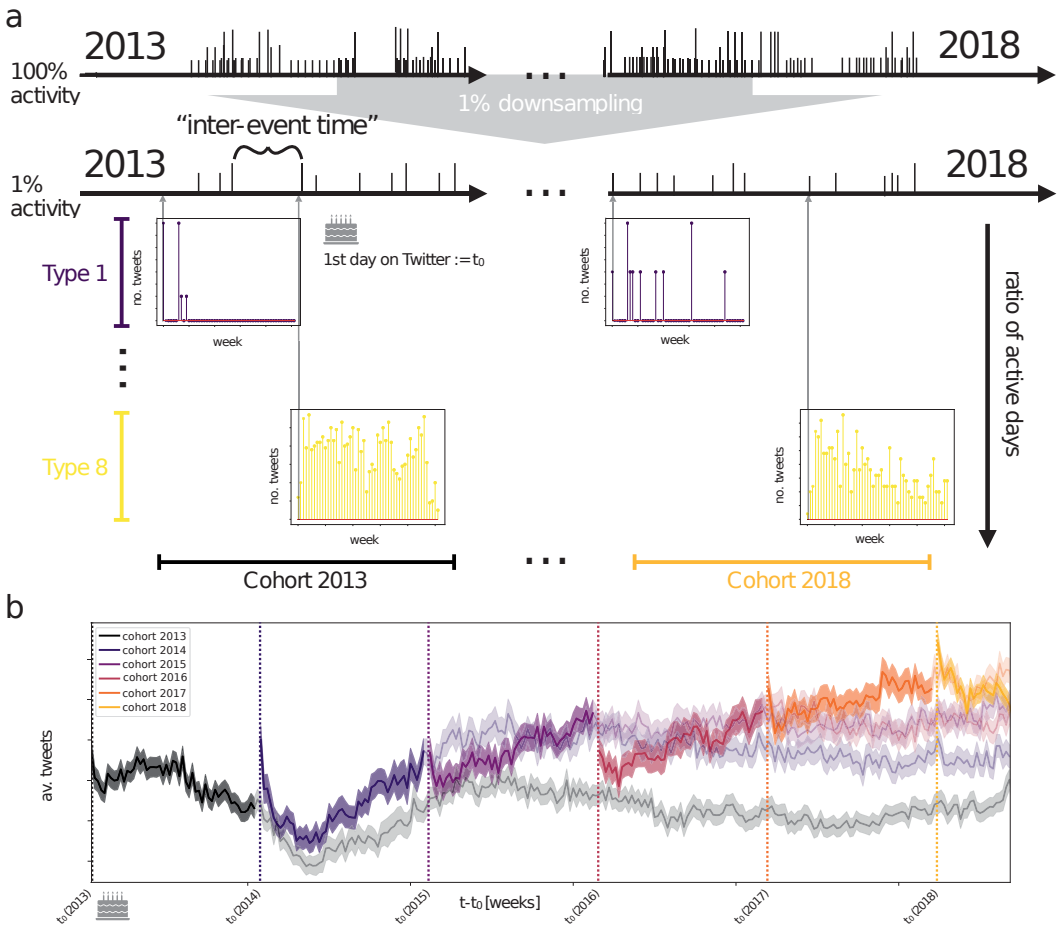

Figure 8.1.: User cohorts and user types in the Twitter data set. (a) Illustration of 1% downsampling from the original tweet trajectory of a single user between 2012 and 2019. Illustration of the difference between user types (vertical axis) and user cohorts (horizontal axis). User types are colored purple to yellow and are sorted by activity, user cohorts are colored black to orange and differ in the year of the first appearance in the data set. (b) Aligned weekly tweets per cohort in the study period. The respective first year is highlighted. Variance is indicated by shaded areas.

8.3. Temporal evolution of user cohorts

To access the potentially different activity of users who have entered Twitter at different points of time, we split the user sample from April 2019 into different cohorts, each depending on the time they first appear in our data set as an active user (sending a tweet or retweet). Subsequently, we analyze the mean tweeting behavior of each user cohort. We show an illustration of the construction of the cohorts in Fig. 8.1a, horizontal axis.

To get an intuitive understanding of how the users in the different user cohorts engage with Twitter, we compute the number of tweets per week for each user. To enable a meaningful comparison of users within a single cohort, as well as between the cohorts, we map all individual weekly tweeting trajectories of each cohort on the first week of the respective year. Thereby, we align all trajectories of each cohort irrespective of their actual starting date. Additionally, we exclude the first tweet, as otherwise all cohorts have elevated activity in their first week (by defining the starting date as the date of the first recorded tweet, all users post a tweet in their first week). The mean of the cohort-wise tweeting trajectory is shown in Fig.8.1b.

Before we discuss the cohort tweeting trajectories, note that we only show the results for the cohorts starting between 2013 and 2018. Users in the cohort from 2012 first appear in 2012 and might have already joined Twitter in 2011 or earlier. Therefore, we interpret the absence of a user in the whole year 2012 as a proxy that this user has not entered Twitter before our study period. In addition, we ignore all users which first appear after April 2018 as we do not have one consecutive year of tweeting for those users.

Investigating the cohort's mean tweeting trajectories (Fig.8.1b), we observe that from year to year, cohorts end up on an increasing level of activity (offset at the end of the study period). Although this comparison is based on tweeting activity of differently "aged" users, this already indicates the different levels of activity of the cohorts. By examining the activity in the first year of each cohort, we additionally observe a trend of increasing activity, especially for the last cohorts. This is an indication that users who have entered Twitter more recently tend to get engaged with Twitter more rapidly and, thus, tweet more frequently from the early beginning.

8.4. Temporal evolution of user types

Another natural way to subdivide users into distinct groups is to split the users by means of activity.

To investigate how the activity is distributed among the users, we first compute the ratio between the days at which we record at least a single tweet of a user and all days between the respective starting date and the end of the study period. Thus, the activity ratio is bounded by $[0, 1]$ and a proxy for the activity of a single user. Using unsupervised k-means clustering and the explained activity criterion, we identify 8

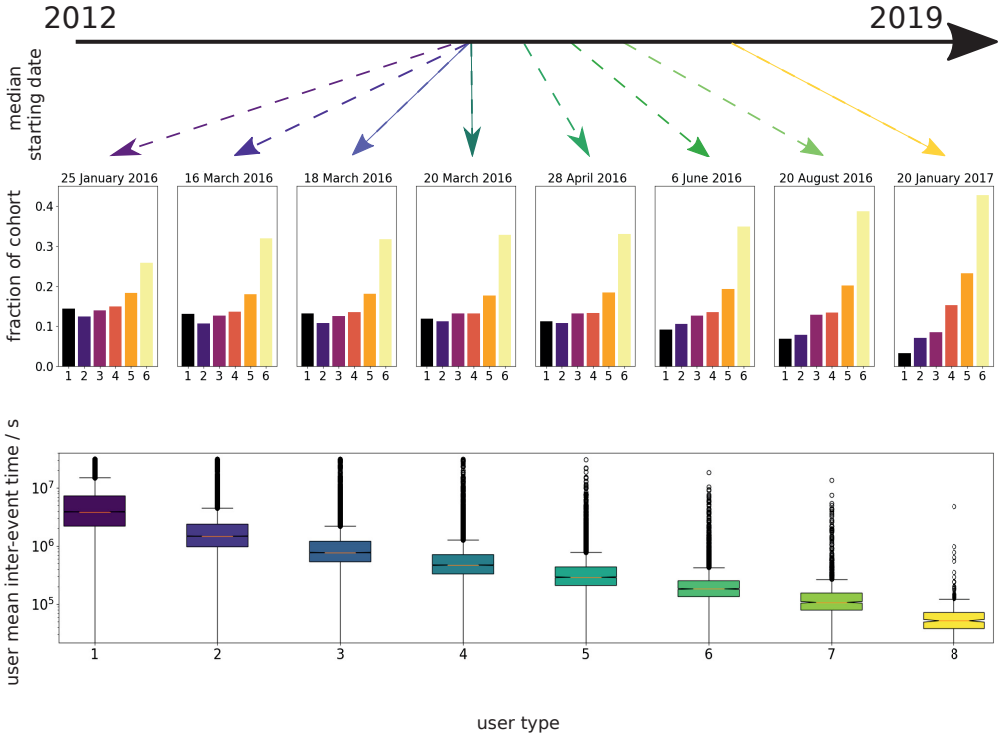

Figure 8.2.: Composition and interevent time of user types. Top: median starting date of each user type indicated by arrow to time axis. (a) Histrograms showing the fraction of users in each cohort for each cluster. (b) User-mean interevent time in the first year for users in the different clusters.

user types with different tweeting activity. We show the schematic difference between user types in Fig. 8.1a. Users of the same user type are colored accordingly.

To connect the different user types with the previous analysis of the cohorts, we subsequently compute the fraction of each cohort for each user type. We show the fractions as histograms in Fig. 8.2 where each histogram (for each user type) sums up to 1. Note, that the last cohort is the largest and the most active user type is the smallest which is not shown in Fig. 8.1. Motivated by the activity offset between the cohorts, we compute the median starting date on Twitter for each user type. We find a one-to-one correspondence between median starting date and the activity of the user type (indicated by the arrows to the time axis, top of Fig. 8.2). This is additionally implied by the heights of cohort fractions in the histogram: we observe that an increasing fraction of recent users ends up as more active user types (decreasing height for cohort 1, increasing height for cohort 6). This also implies that user type 8 grows the quickest relative to its size.

To finally investigate the, by definition, dissimilar user activity of the user types, we compute the mean waiting time between two recorded tweets of each user. The bottom panel of Fig. 8.2 shows the corresponding user-type-wise box plot. As we

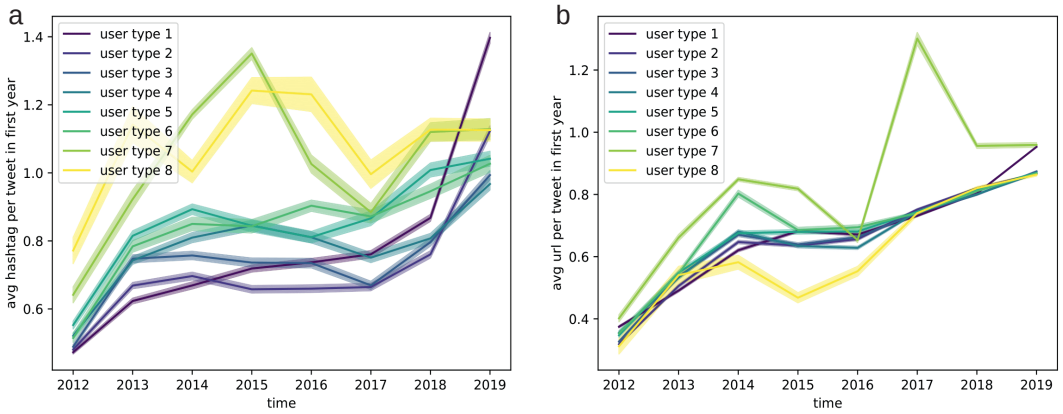

Figure 8.3.: Average (a) hashtags and (b) URLs per tweet for each user type over time. Only users that have posted at least one hashtag or URL are considered.

have obtained the user types by measuring the activity, the systematic differences (note the logarithmic axis in the bottom panel of Fig. 8.2) are not surprising. The striking finding is the one-to-one correspondence between the mean age ranking and the activity ranking. We assume that this finding further supports the hypothesis of increasing individual activity of Twitter users.

With user types separated into different groups, we have obtained a user partition which allows for long term study. Although each user type grows in size with time, we can compare the different user types over the whole study period to systematically investigate different behavior of user types.

Accordingly, we analyze the content of the tweets which is shared by users of a different type. We compute the average number of URLs and hashtags which are shared by users who have at least once shared an URL or a hashtag on Twitter. In Fig. 8.3a, we show the result for the different user types and the evolution of the hashtag usage over the study period. As a general trend, we observe that the average number of hashtags has increased from ≈ 0.6 hashtags per tweet to ≈ 1 hashtag per tweet. Additionally, more active user types tend to use more hashtags per tweet.

In addition to hashtags, sharing URLs is an effective way to share opinions and information with other users. Utilizing the same framework as for the hashtags, we compare the URL use of the different user types in Fig. 8.3b. In contrast to the evolution of hashtag use, Fig. 8.3b shows no coherent separation between the user types. While hashtag use has been found to depend on the activity of the users the use of URLs seems to be independent. Although the user types exhibit more or less indistinguishable numbers of URLs per post for all years, they agree on the increasing trend which supports the hypothesis of users sharing more content and connecting their posts to more topics.

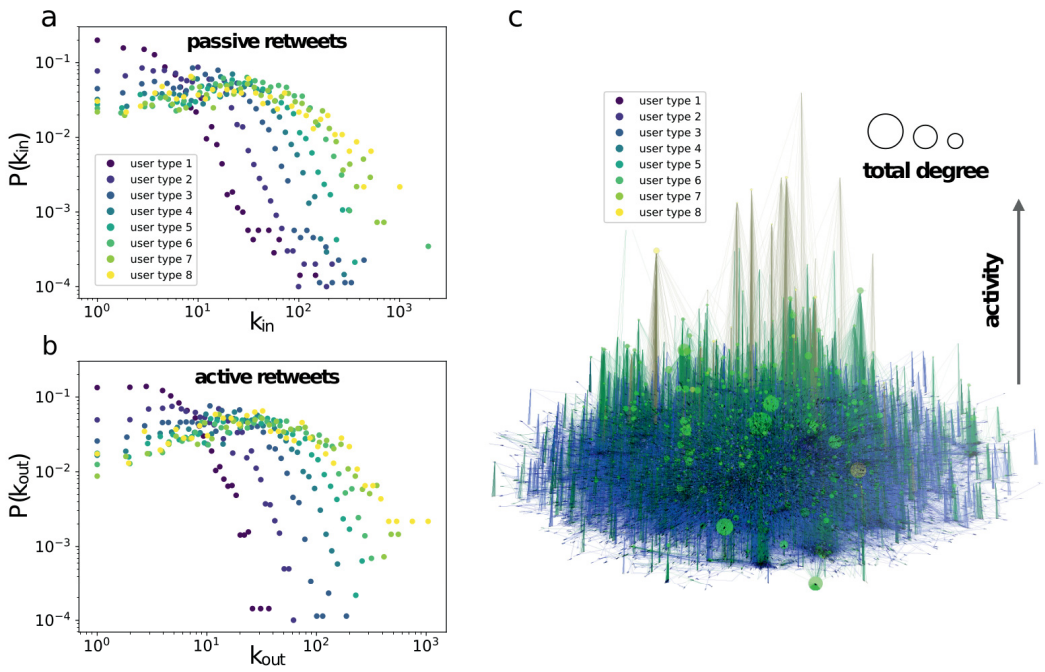

Figure 8.4.: Properties of retweet interactions. (a) In-degree and b) out-degree distribution for each user type of the retweet interaction network. (c) Illustration of the retweet network with using gephi [51]. Nodes are colored by the user type. Diameter of the nodes varies according to the total degree.

8.5. User types in the retweet network

To investigate how tweeting behavior is related to the interaction on social platforms, we study how activity translates to the position in the interaction network. Based on the retweets in our data set, we introduce a directed edge pointing from the actively retweeting user towards the user that is passively being retweeted and, thus, construct a weighted interaction network. In other words, we update the elements w_{ij} of the weight matrix $\boldsymbol{W}$, by $w_{ij}^{new} = w_{ij} + 1$ for each retweet of user i retweeting user j. The results of the network analysis are shown in Fig. 8.4.

Figures 8.4a,b depict the in- and out-degree (Eqs. 2.6 and 2.7, replace a_{ij} with w_{ij}) distributions which relate to the passive and active tweeting behavior of the users. This basic analysis highlights differences between the user types and reveals that users with a higher tweeting activity are more central in the network. This is reflected by a broader distribution of the in- and out-degree. As the out-degree is directly connected to the tweeting activity of a single user, we can explain the clearer separation between the user types by their distinct activity. The difference in the in-degree distribution reveals an interesting feature and hints already to the overall network structure. More active users do not only tweet and retweet more but also get retweeted more often and are, therefore, expected to be represented by

more central nodes in the interaction network. Figure 8.4c shows a representation of the interaction network with two dimensions of differentiation. Firstly, nodes with a higher degree are marked with a larger diameter, and, secondly nodes are vertically aligned by means of their tweeting activity. The color code represents their user-type affiliation. This representation highlights that users from the more active user type (which are on average "younger") are more central (larger diameter) in the network and inactive user types build up the periphery.

In the previous sections, we have shown that active users show a qualitatively different behavior, use more hashtags, and, thus, increase the density of information and more likely have joined Twitter in recent years. Besides, we have observed that the response of the network, the number of passive retweets, aligns with the number of active tweets.

To investigate how different levels of activity shape and influence topical persistence, we further elaborate on the used hashtags of users in the following.

8.6. Individual content horizon

As a first step, we examine how the content in tweets and passive retweets aligns for different delay times (see Fig. 8.5a). This illustrates whether and how topics in the tweets of a user relate to the topics that have been retweeted by other users in the past or will be retweeted in the future. Note that we have normalized the correlation using $U_{ij}(0)$, which we define as the mean of $U_{ij}(-1 \text{ week})$ and $U_{ij}(1 \text{ week})$. As $U_{ij}(0)$ refers, by definition, to the maximal overlap of 1 which is one order of magnitude larger than the other correlation scores, we have removed this value. Without normalization, the correlation is higher for low activity user types (not shown).

In general, less active user types align longer with their passively retweeted content (indicated by larger correlation for increased delay). This can be related to a quicker shift of topics for the high active users or possibly implies that highly active users do not align their content with their own peers. This rapid transition towards sharing new content can occur intentionally with the goal to *influence* the peers. To get a clearer picture of that, we study the topical autocorrelation of the own tweets which is shown in Fig. 8.5b which is normalized as described above. This normalization enables us to compare the slopes of the user type autocorrelation functions. We do not show the systematic offset between the autocorrelation functions. As in Fig. 8.5a, less active user types exhibit larger topical autocorrelation.

We suspect that the findings in Fig. 8.5 point to multiple features of the communication among user types. First, less active user types have a large topical persistence (flat autocorrelation) and, thus, share similar content for a longer period. This might imply that less active users spend more time with distinct topics than more active users. Second, for the more active users, we observe decreasing topical overlap with increasing delay. Whereas the general offset between the autocorrelations (which we have not shown) is likely being related to the fewer hashtags posted each week by less active users, the heterogeneous decay of the autocorrelation functions indicates

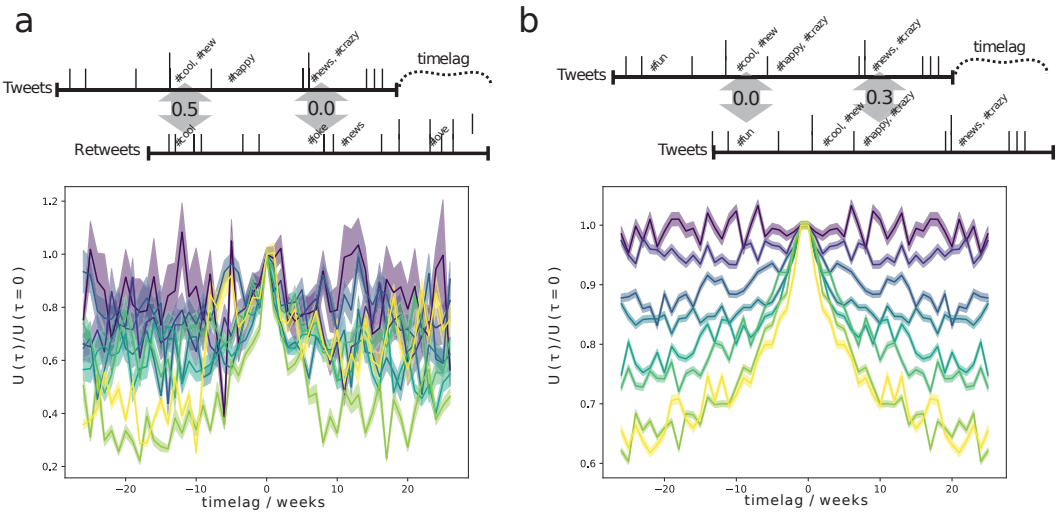

Figure 8.5.: Normalized topical correlation between tweets and passive retweets (a) and own tweets (autocorrelation) (b) for varying delay. On top of both figures we schematically illustrate measuring topical correlation.

that more active user types spend less time with the same topic(s) before drawing their attention to new trends. We assume that this is possibly caused by a limit of the individual content horizon: with enhanced activity, users get exposed to more topics and, to maintain an overview on the trending subjects, reduce the attention towards older topics after some time. This effect can additionally be amplified by the construction of Twitter as a so-called *feed*. New posts are presented in a hierarchical order and older posts do not appear in the feed after some time (without extensive scrolling). This draws attention to more recent topics. As more active users have a more crowded feed, the organization of the Twitter application might facilitate this effect.

Individual content horizon and increasing activity

In the first sections of this chapter, we have shown that more recent users tend to use Twitter more actively. This rising activity has been characterized by increasing numbers of tweets per week and decreasing inter-event times. Such a development together with the observation of a shrinking content horizon can lead to multiple potentially undesired effects. Among others, shorter individual attention towards so-called *fake news* can undermine efforts to correct false statements, might facilitate the need for simplification of complex problems, and possibly impedes a focus on long-term challenges. As we assume that our finding can be confirmed by studying other data sets (already mentioned empirical evidence for collective acceleration [147] or trend of decreasing desktop time on news webpages [176]), we want to emphasize that the possibly accelerating nature of collective and individual interactions and the corresponding reduced content horizon have changed and will change how people

interact online. Future studies covering more data sets, examining potential qualitative shifts or social tipping points, and suggesting constructive interventions such as frictions are crucial to understand how we, as societies, should cope with these evolvements.

8.7. Summary

In this chapter of the thesis, we have investigated user behavior on the social media platform Twitter in a longitudinal study. By analyzing tweets from 600,000 Twitter users, we have studied the frequency and content of online interactions during the time between January 2012 and June 2019.

Initially, we have addressed how the Twitter environment, which has changed over the last decade, has succeeded in engaging the users to actively use social media. Therefore, we have split up the users into cohorts with a mutual starting year. For each of the cohorts, we have analyzed the mean weekly tweeting activity. This analysis has shown that the level of activity has not only increased comparing the first year of each cohort but has also led to a separation of cohorts by means of activity on the long run.

Subsequently, we have investigated different user types. By applying an activity-based clustering scheme, we have split the users into 8 user types with distinct activity. To link these user types with the cohorts, we have computed the relative fraction of users of the different cohorts in each user type. This has revealed that more recent users dominate the more active user types. We have, therefore, concluded that the more active user types grow quicker relative to their size than the less active user types. We have further illustrated this finding by computing the median starting day of each user type. This has revealed the one-to-one correspondence of more recent starting days and shorter user-mean inter-event times. Along with the study of the tweeting frequency of the user types, we have shown that increased activity coincides with more frequent use of hashtags. Furthermore, the numbers of URLs and hashtags per tweet have almost doubled in the last eight years.

Thereafter, we have investigated how the different activity of users shapes the topological neighborhood in the social network of Twitter. We have set up a network by using the retweets among the users. By analyzing the out-degree (referring to the active retweeting of a user) and the in-degree (referring to the passive retweets of a user), we have found that more active user types are more central with respect to their in- *and* out-degree. This has proven that the attention of the own followers is not only raised by quality but also quantity of tweets.

By finally investigating the topical correlation of the tweets, we have shown that more active users switch topics more quickly and, therefore, exhibit a shorter individual content horizon.

We conclude that interactions between single users on Twitter have accelerated over time. We suspect that this quantitive different behavior can lead to qualitative shifts of collective online dynamics especially in combination with the described

co-occurrence of increasing content per tweet and prominent positions of highly active users in the Twitter network. In a world where public discourses are at least partly discussed on social media, we assume that this observation has to be evaluated critically and constructive interventions should be considered to cope with the ongoing developments.

Chapter 9.

Conclusions and Outlook

The main goal of this thesis has been to promote data-driven investigations of dynamic processes using complex networks employing a twofold approach: on the one side, we have aimed for improving the theoretical framework by introducing novel measures for networks embedded in spherical geometry. On the other side, we have tailored diverse specialized tools and conducted several studies in the fields of Climatology and Computational Social Science utilizing complex networks. For that, we have separated the thesis into three parts. In the following, we review the main findings of these three parts before we close the thesis with an outlook, elaborating on important tasks for future research.

9.1. Conclusions from part 1 - Methodological approaches to study complex networks

In the first part of the thesis, we have introduced complex networks as a tool to study complex systems and have reviewed the theoretical foundations as pre-requisites for understanding the analyses performed. We have also proposed a methodological framework to quantify edge directionality in complex spherical networks in the last chapter (chapter 3).

To extend the toolbox of measures suited to study spatially embedded networks, we have recalled the concept of edge anisotropy which has been proposed for flow networks in two dimensions. Following the rationale of this previous work, we have introduced the mean edge direction and the mean angle as additional network measures. Motivated by the study of global real-world networks, we have introduced their representation for networks embedded in a two-dimensional spherical surface representing planet Earth. Next, we have studied a generic bias that can arise from heterogeneous node placement and have suggested a suitable correction scheme. With the analysis framework in place, we have finally illustrated the broad applicability of edge directionality measures by studying different climate networks, a trade network, and an air transportation network. In the climate networks, we have identified the edges of the circulation cells as well as structures related to the El Niño–Southern Oscillation and the Atlantic Niño. In the trade network, we have highlighted the influence of the EU on the global import network as well as classified countries by their alignment of import and export directions. Finally, we have recognized regions

of distinct air transport in the respective network. All these analyses have been based on edge directionality measures only.

9.2. Conclusions from part 2 - Complex network in Climatology

In the second part, we have studied the Earth's climate system employing functional climate networks.

In the first chapter (chapter 4), we have comprehensively compared different strategies to construct climate networks from extreme event time series. In particular, we have utilized Event Coincidence Analysis and Event Synchronization to investigate synchronous heavy rainfall related to the South American Monsoon System. In their original definition, both methods lead to distinct networks although they predominantly only differ regarding their exact definition of event coincidence. By considering the pairing coefficient which quantifies event clustering in event time series, we confirm a generic bias of Event Synchronization. Subsequently, we have analyzed the results from the application of a corresponding correction scheme, which does not only account for the existing bias but can also be motivated physically by interpreting event clusters as long persistent events. Considering the latter, we have shown that the application of the correction scheme can also be meaningful for Event Coincidence Analysis which is not prone to the systematic clustering bias. To further illustrate the systematic difference between the local definition of the coincidence interval of the Event Synchronization and the global setting of the coincidence interval in the Event Coincidence Analysis, we have conducted an additional case study, where we have highlighted the advantages of both methods: whereas the global setting of the delay and coincidence interval in the Event Coincidence Analysis allows for tracking rainfall event cascades in great detail, the dynamic character of Event Synchronization enables us to study event cascades as a whole without setting parameters.

In the second chapter (chapter 5), we have utilized Event Coincidence Analysis to study heavy rainfall events associated with the Baiu frontal system. Throughout our analysis, we have identified a spatially separated double band of synchronous heavy rainfall. The formation and withdrawal of this double band seem to be intimately linked to the onset and withdrawal of the Baiu front. Starting by studying monthly network patterns, we have validated our observation by investigating the community structure during the period of high inter-band-synchrony, the development of the total cross-degree between the double band over time, and by studying rainfall composites of days with large event numbers in both regions. Finally, we have studied additional meteorological variables such as geopotential height, winds in different altitude, and relative vorticity. This has led to the conclusion, that the double band is related to the interplay of the North Pacific Subtropical High and the South Asian Anticyclone, which have been considered independent in previous studies.

To complement these two real-world climate network analyses, we have examined model output from the global circulation models within the TRACMIP in chapter 6.

By clustering the respective 14 comprehensive aquaplanet model runs utilizing their zonal mean network measures, we have classified the models into groups with distinct ITCZ dynamics. This has enabled us to relate specific structures and features of the zonal mean network measures to the time mean position of the ITCZ. In addition to the analysis of the global network, we have re-run the clustering framework for scenarios with selected linkage patterns (e.g. excluded extratropic-extratropic connections) and have shown that we need to consider tropical-tropical as well as tropical-extratropical links to obtain meaningful clusters of climate models. With this work, we have confirmed the global character of the ITCZ and its related dynamics and proven that SST variability (in contrast to the time-mean tropical SST difference) and the time-mean ITCZ position are intimately linked.

9.3. Conclusions from part 3 - Complex networks in Computational Social Science

In chapter 7, we have investigated the temporal change of interactions on the fashion blog `lookbook.nu` in the year 2015. Initially, we have constructed weekly networks based on aggregated co-occurrences of hashtags. Analyzing these static networks, we have not only shown that hashtags are hierarchically organized but also found that hashtags on this fashion platform mainly form two different classes of communities. Whereas the first has been characterized by a hub at the center of communities and adjacent descriptive hashtags with few other connections, the second has been highly interconnected and exhibits an elevated community-mean clustering coefficient. To study the dynamic temporal change of the communities, we have introduced a novel matching scheme which we have based on a temporal extension of the Jaccard index, the so-called memory weights. The method allows for automatic relabeling of community allocation considering different orders of memory. After basic stability testing, we have applied the method to study the temporal emergence of topics on `lookbook.nu`. We have found that topics and trends which we have identified with the different communities in the hashtag co-occurrence network change over time. In particular, we have revealed the seasonal change of topic popularity and discovered stable distinct fashion genres. Finally, we have conducted a comparison of the popularity trajectories around their respective peaks and have shown that popularity is gained and lost in a surprisingly similar fashion.

In the last chapter of the thesis (chapter 8), we have studied the long-term change of individual user behavior on the social platform Twitter. For this study, we have had access to a large data set of interactions on Twitter covering the period from January 2012 to June 2019. Initially, we have grouped users into cohorts with mutual starting year and have shown that users in more recent cohorts tend to be more active which we attribute to a shift in the cohort mean activity. To enable the long-term study of distinct user groups, we have clustered the users by means of their activity and, thus, have identified distinct user types. The activity of these user types has differed dramatically and has revealed the broad activity range from passive to addictive user

behavior. We have linked the user types to the cohorts by computing the relative cohort fraction of each cohort for the user types. We have shown that the most active user types exhibit a more recent median starting date and that the fraction of more recent users coherently increases with the activity of the user type. Examining the content of the tweets has revealed that all users have significantly increased their hashtags and URLs per tweet indicating the rising information shared on Twitter. Considering that users having joined Twitter more recently use Twitter more actively and that more active users exhibit a smaller individual content horizon as well as increase their hashtags and URLs per tweet, we have presented empirical evidence for a social acceleration on an individual user level. This individual acceleration might lead to a collective acceleration as active users are located at more central positions in the interaction network.

9.4. Outlook

As illustrated by the multiple different applications presented in this thesis, network approaches to study complex systems have a great potential to distill information from the data that is collected, regardless of whether we aim for analyzing sea surface temperature variability or interactions on social media. Therefore, Network Science and associated approaches are state-of-the-art tools in many different fields of study. Before we review the challenges and possible future tasks of the two application fields (Climatology and Computational Social Science), we want to stress the potentials of interdisciplinary research. Our work was inspired and supported by findings from completely different fields. Methods such as Event Synchronization have originally been developed for Neuroscience and network measures such as edge anisotropy have been developed to study the directionality of flows. Utilizing concepts from other fields of applications often opens the doors for completely new views on many tasks. This is one of the most striking advantages of Network Science. To set already developed methods in a new context should always be considered as a natural strategy in our field. Despite this general focus on interdisciplinarity, many concrete research questions are arising from the results in this thesis.

Utilizing complex networks to study the Earth's climate system has led to great advances in Climatology. We believe that there is potential for other network-based precise onset prediction schemes besides the successful onset scheme for the Indian Summer Monsoon. In this thesis, we have presented the first step towards such a scheme for the East Asian Summer Monsoon. In this particular case, further investigations on the characteristics of the newly observed double band are of utmost importance to clarify its relation to the Baiu onset.

As a general problem, we have identified the resolution limit of temporal network analysis using the common approach of sliding windows as one conceptual limitation. As most of the recent network science-related findings in Climatology are based on Event Synchronization or Event Coincidence Analysis, sufficiently long event time series are crucial ingredients for such analyses. An alternative novel method to

construct such networks could possibly set the stage for even more accurate prediction schemes.

To our knowledge no network-based study has yet been conducted on monsoonal rainfall in Africa and only one recent study on the Australian Monsoon System [177]. To better understand extreme events and their precursors in general, a comprehensive comparison between the singular studies might contribute to a generalized picture of the climatology of extreme events.

In addition to the long-term goal of a generalized prediction scheme for the formation of major weather systems and associated extreme events, the investigation of tipping points is an increasingly important topic in Earth System Science. While most of the recent studies focus on conceptual work linking a small number of tipping elements, we believe that Network Science can serve as a toolbox to investigate regime shifts on a greater scale. Describing tipping elements by groups of nodes or communities in a network can eventually point out the most critical couplings.

Finally, we have used complex networks to examine idealized model output data. Working together with experts in the field of modeling has brought together two different approaches to studying ongoing climate change. Making use of network science methods to analyze, classify, and validate climate model output can possibly be used to advance the precision climate change projections. In this scope we emphasize that disentangling the relation between the time-mean ITCZ position and the SST variability can possibly further clarify why global circulation models often fail to correctly reproduce ITCZ dynamics.

In the context of Social Science, complex networks have been utilized since computational power allows for the analysis of the ever-rising amount of collected data. In our opinion, merging Complex System Science with Social Sciences is an important and promising collaboration. To constantly evaluate the impacts of modern communication, theoretical modeling and data analysis are crucial. Nowadays, researchers can base their findings on data with a temporal resolution of seconds. As the content of the data is expanding every year, completely new analysis frameworks enable us to characterize human interactions with fundamental laws. Specifically, modeling the rise and fall of social networks, the formation of echo chambers within these networks, and information diffusion as well as opinion formation are well-established research fields. Further research in this direction can guide policymakers on the super-national level to establish frameworks in which the internet and social media can transform such that societies profit from online interaction. Our effort to develop a tool for tracking the popularity of topics on social media is one step towards a clearer picture of online popularity in general.

Simultaneously, the observed acceleration of online discourses and the decreasing timespan of public attention challenge democracy and the organization of our society. We outline the study of intervention strategies such as friction (decelerate the interaction and posting frequency) and individual choices for information presentation (enable users to understand why they see which posts) as topics for further investigations. Summarizing, studying modern ways of interaction in empirical studies contributes to the basic understanding of how we, as humans, communicate online.

We do not only need such information to set up useful models but also to anticipate eventual consequences of the environment design. We, therefore, emphasize the importance of validating our finding regarding the acceleration of individual user interactions in data sets from other social media platforms.

Appendix

Appendix A.

Analytical expression for spherical course angles

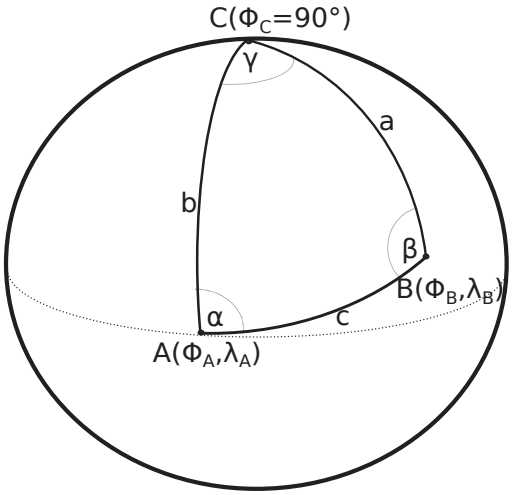

Figure A.1.: Sketch of a spherical triangle ΔABC. We identify point C with the North Pole ($\phi_C = 90°$).

To derive the analytic expression for the course angle (Eq. 3.6), we start by considering an arbitrary spherical triangle which we illustrate in Fig.A.1. The triangle we consider is characterized by the three corners (labeled A,B and C) and the sides with length a,b and c. In line with common notation, the side with length a is opposite to corner A and the angle α is measured at corner A between the sides with length b and c. All other labels (angle β and γ and sides with length b and c) are located accordingly. Note that we specify length in radiants.

For such a spherical triangle, we recall the spherical law of cosine (cosine rule for sides on a sphere) as

$$\cos a = \cos b \cos c + \sin b \sin c \cos \alpha$$

$$\Leftrightarrow \qquad \cos \alpha = \frac{\cos a - \cos b \cos c}{\sin b \sin c}. \tag{A.1}$$

Note that lengths are measured as angular distance in radiants (and correspondingly need to be multiplied by the Earth's radius to match with real-world networks).

In our application we quantify the angle α between the side with length c and the true northern direction for a triangle as shown in Fig. A.1. Therefore, we identify the corner C with the North Pole. This simplifies the caclulations as we measure the angular distance between between A and C (and B and C, respectively) in terms of the latitudes. As the latitude increases from the equator to the poles, we set $\cos a = \sin \phi_B$, $\cos b = \sin \phi_A$, and $\sin b = \cos \phi_A$. Replacing the according terms in Eq. A.1 leads to

$$\cos \alpha = \frac{\sin \phi_B - \sin \phi_A \cos c}{\cos \phi_A \sin c} = \frac{\sin \phi_B - \sin \phi_A \cos c}{\cos \phi_A \sqrt{1 - \cos^2 c}}. \tag{A.2}$$

where we ignore the second possible (negative) solution of $\sin c = \pm\sqrt{1 - \cos^2 c}$ for the second identity. This (negative) solution refers to the supplement angle which is not of interest in our considerations.

In addition, we can express the angular distance between A and B by considering the corresponding scalar product in spherical cooridinates. Simple calculations show

$$\cos c = \sin \phi_A \sin \phi_B + \cos \phi_A \cos \phi_B \cos(\lambda_B - \lambda_A). \tag{A.3}$$

Inserting Eq. (A.3) into the numerator of Eq. (A.2) leads to

$$\sin \phi_B - \sin \phi_A \left(\sin \phi_A \sin \phi_B + \cos \phi_A \cos \phi_B \cos(\lambda_B - \lambda_A)\right) \tag{A.4}$$
$$\Leftrightarrow \sin \phi_B - \sin^2 \phi_A \sin \phi_B - \sin \phi_A \cos \phi_A \cos \phi_B \cos(\lambda_B - \lambda_A) \tag{A.5}$$
$$\Leftrightarrow \sin \phi_B (1 - \sin^2 \phi_A) - \sin \phi_A \cos \phi_A \cos \phi_B \cos(\lambda_B - \lambda_A). \tag{A.6}$$

Using the identity $(1 - \sin^2 \phi_A) = \cos^2 \phi_A$ we can further simplify the numerator

$$\Leftrightarrow \sin \phi_B \cos^2 \phi_A - \sin \phi_A \cos \phi_A \cos \phi_B \cos(\lambda_B - \lambda_A) \tag{A.7}$$
$$\Leftrightarrow \cos \phi_A \left(\sin \phi_B \cos \phi_A - \sin \phi_A \cos \phi_B \cos(\lambda_B - \lambda_A)\right). \tag{A.8}$$

As $\phi_A \neq 90°$ we can now cancel $\cos \phi_A$ in Eq. (A.2). To obtain Eq. (3.6) in chapter 3 we finally have to insert Eq. (A.3) into the denominator and replace $A \to i, B \to j$ to match the stated notation.

Appendix B.

Extension of the method intercomparison between ES and ECA

Introduction

In this chapter, we outline further investigations related the method intercomparison presented in chapter 4 section 4.5. In particular, we elaborate on differences regarding the spatial network measure pattern and network measure correlations between networks based on ES and ECA. For both, we examine their relation to varying the link density ρ (see Eq. 2.1).

The results presented and the figures shown in this chapter are based on publication P_3 *Wolf, F., and Donner, R.V. (submitted to European Physical Journal - Special Topics). Spatial organization of functional climate network characteristics describing event synchrony of heavy precipitation.* We thank EPJ for the kind permission.

Motivation and study setup

Many studies have utilized event synchrony measures to study certain features of the Earth's climate system. In table B.1, we show an overview of some previous work. We do not claim that this list is complete by any means. A large fraction of the investigations employed a fixed link density between $\rho = 2\%$ and $\rho = 10\%$. Next to this systematic difference, there have been examined various network measures. In chapter 4, we have solely focussed on the information encoded in the degree pattern. Here, we illustrate the impacts of modifications regarding the analysis setup on functional climate network analysis. Specifically, we first show spatial network measure patterns obtained by modifying the link density. Second, we examine single network measure correlations as in Figure 4.5. Finally, we present multivariable correlations and comment on their potential to complement climate network analyses and to illustrate event synchrony method dissimilarities. As in chapter 4, we study network characteristics for the SAMS setting an upper limit for the dynamic coincidence interval of the ES ($\tau_{max} = 3$ days) (see Eq. 2.17) and choosing $\Delta T = 3$ (see Eq. 2.22) as the size of the global coincidence interval of the ECA. For the analysis, we again employ the declustering scheme for both ES and ECA as well as the average symmetrization (see Eq. 2.26) of the ECA.

Table B.1.: Previous publications employing event synchrony measures to study functional connectivity in the Earth's climate system based on rainfall data (exeption: Boers et al. [55] employed moisture divergence). We have used the following notations for the network measures: k degree, c local clustering coefficient, v closeness, b betweenness, Δk network divergence, $\mathcal{V}$ vulnerability, $\mathcal{L}$ long distance directedness, d average link distance, Δs node strength divergence, $\mathcal{R}$ regional connectivity, $\mathcal{M}$ MSES value, $\mathcal{P}$ participation coefficient, Bri bridgeness, BIL degree and influence of line, Q molularity. For a definition of the different characteristics which we have not specified in the theoretical foundations (chapter2), we refer to the corresponding publications.

authors	year	region	time covered	link density	network measures
Malik et al. [95]	2010	ISM	$1961-2004$	–	–
Malik et al. [34]	2012	ISM	$1951-2007$	5%	$k, c, v, b, \Delta k, \mathcal{V}$
Boers et al. [29]	2013	SAMS	$1998-2012$	–	$k, \mathcal{L}, d, v, b, \mathcal{L}$
Boers et al. [5]	2014	SAMS	$1998-2012$	–	Δs
Boers et al. [97]	2014	SAMS	$1998-2012$	2%	$k, c, d, b, \mathcal{L}$
Boers et al. [55]	2014	SAMS	$1979-2010$	5%	c
Stolbova et al. [70]	2014	ISM	$1951-2012$	5%	k, b, d
Su-Hong et al. [178]	2014	EASM	$1951-2007$	5%	$d, \mathcal{V}$
Boers et al. [96]	2015	SAMS	$1998-2012$	–	$\mathcal{R}$
Boers et al. [98]	2015	SAMS	$1979-2013$	2%	$k, c, b, \mathcal{R}$
Agarwal et al.[76]	2017	Germany	$1901-2010$	–	$\mathcal{M}$
Agarwal et al.[179]	2018	ISM	$1901-2013$	–	$k, \mathcal{P}$
Ozturk et al. [180]	2018	East Asia	$1998-2015$	5%	Δk
Agarwal et al.[181]	2019	Germany	$1901-2010$	–	k, c
Boers et al. [24]	2019	worldwide	$1998-2016$	–	k
Kurths et al. [182]	2019	ISM	$1951-2013$	5%	$\mathcal{P}, Q$
Ozturk et al. [75]	2019	EASM	$1998-2015$	5%-10%	k, d, v, c
Agarwal et al.[183]	2020	Germany	$1901-2010$	–	$k, b, \mathrm{Bri}, \mathrm{BIL}$
Cheung et al. [177]	2020	AMS	$1998-2015$	5%	k, c
Wolf et al. P_2	2020	SAMS	$1998-2015$	2%	k
Wolf et al. P_4	2020	EASM	$1998-2018$	5%	k, d

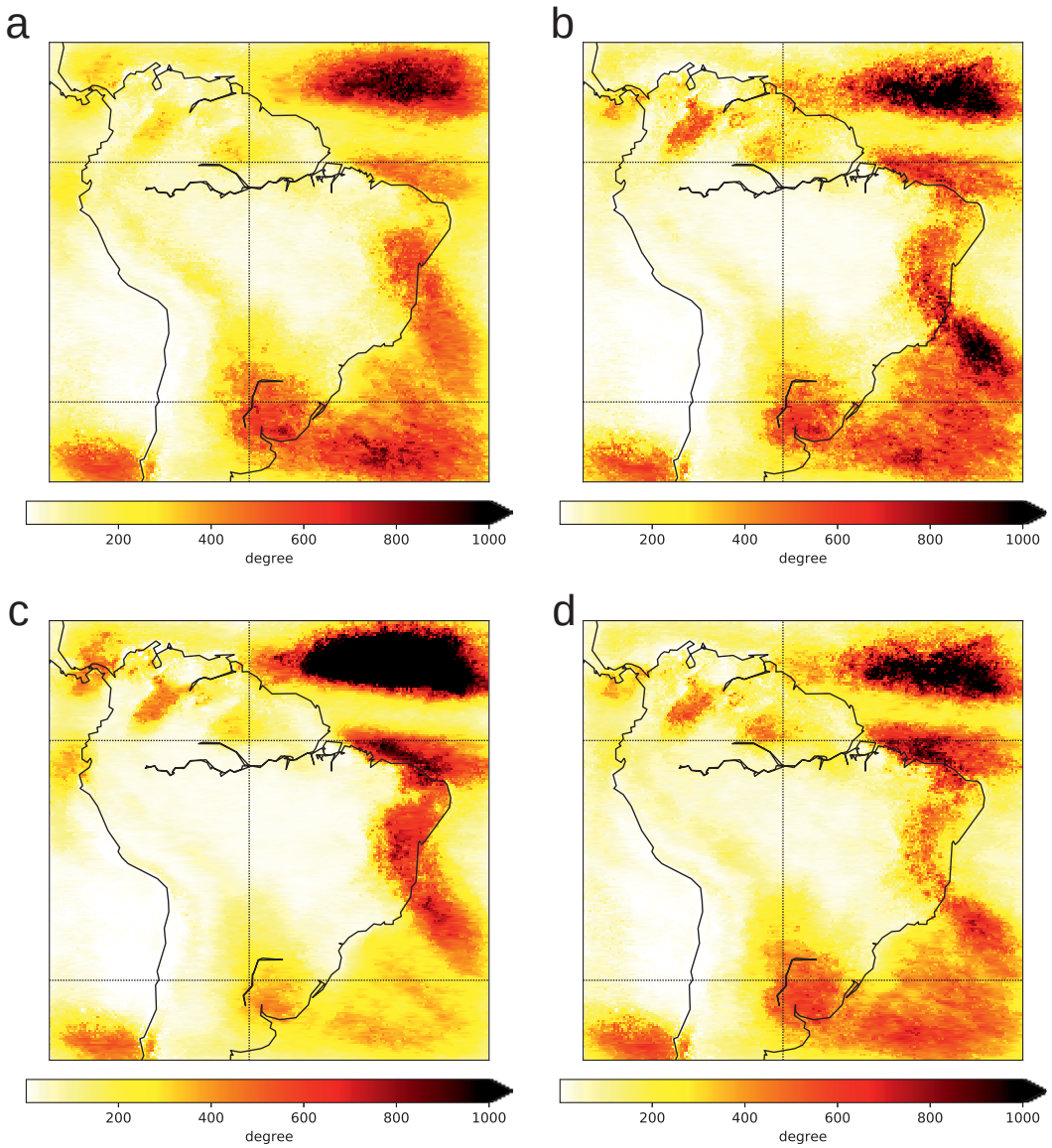

Figure B.1.: Node degree patterns of the functional climate network representations of heavy rainfall events based on (a,b) ES and (c,d) ECA without (a,c) and with (b,d) incorporating the declustering scheme. All networks exhibit a link density of $\rho = 0.5\%$.

Spatial network measures

Node degree

To tie in with the analysis in part two of the thesis, we first examine the degree patterns. In contrast to the previous analysis where we have utilized a link density of $\rho = 2\%$, we here show the node degree pattern of the SAMS for the link density of $\rho = 0.5\%$ and $\rho = 5\%$ in Figs. B.1 and B.2, respectively. The two cases, therefore, illustrate two scenarios with decreased and increased link density.

In Fig. B.1 we present the degree (see Eq. 2.5) patterns for ES (top) and ECA (bottom) without (left) and with (right) utilizing the correction scheme. As for the degree patterns in the networks with a link distance of 2% (Fig. 4.3), the uncorrected versions of ES and ECA exhibit major differences, whereas the corrected versions appear to have similar features. Again, in the corrected ES-based and ECA-based networks, both the rainfall dipole as well as the ITCZ are characterized by elevated degree values. In turn, we observe a band of high degree north of the ITCZ in the Atlantic Ocean and only slightly elevated degree in SESA in the uncorrected ECA based network. In the uncorrected ES-based network, there is a clear sign of the South American rainfall dipole but the ITCZ is less expressed. For both measures, we observe two patches of an elevated degree in the northern Amazon basin, whereas the moisture pathways along the East Central Andes are hardly visible.

To examine the influence of elevated link distance, we show the degree pattern for a link density of 5% in Fig. B.2. Comparing the degree pattern of the uncorrected versions of ES and ECA (Fig. B.2a,c), we find that the structures in the degree field appear like negatives of each other.

The uncorrected ES-based network is now characterized by spatially extended areas of an elevated degree over continental South America. Specifically, the Southwestern part of the South American rainfall dipole over SESA (along with those regions of the southwestern Atlantic ocean associated with the southern part of the rainfall dipole) and the Andes mountain range as well as the southern Amazon basin display high degree values. As before, the ITCZ and the eastern part of the South American rainfall dipole are less marked. Comparing the pattern to the uncorrected ES-based network in Fig. B.1, there has been a prominent shift in the highlighted regions. To our best knowledge, a similar crossover behavior has not been described in ES-based climate networks before but was reported for the global (scalar) clustering coefficients of global surface air temperature networks established based on classical Pearson correlation [57].

The uncorrected ECA-based climate network still closely resembles the pattern we have observed employing a smaller link density. This includes particularly high degree values along the ITCZ, the northern Amazon basin, and SEBRA whereas the southern part of the South American Rainfall dipole in SESA is marked by an intermediate degree.

Utilizing the declustering scheme again results in a notably more similar degree pattern. As the differences have been much greater in the networks with elevated

link density, we emphasize the impact of the declustering scheme. Although the main features of the SAMS are displayed in both degree pattern (Fig. B.2b,d) there remain certain contrasts. Especially the separation between the two patches related to the South American rainfall dipole and the degree along the eastern flank of the Andes are expressed differently in the degree pattern.

Average link distance

To complement the topological information obtained by analyzing the node degree, we subsequently compute the average link distance (Eq. 2.30). Considering the spatial information of link lengths allows for distinguishing between regional connectivity (e.g., locally increased spatial autocorrelation of the climate variable of interest) or the emergence of large-scale teleconnectivity [57].

The spatial pattern of the average link distance employing a link density of $\rho = 0.5\%$ (see Fig. B.3) resembles the corresponding degree pattern of the corresponding analysis (corrected or uncorrected ES/ECA) with the same link density (compare Fig. B.1 and Fig. B.3). This also holds for an enhanced link density of $\rho = 5\%$ (not shown), where link distance patterns highlight more or less the key features of the SAMS for all analysis setups. As a contrast, the patterns based on the uncorrected synchrony measures differ substantially. The pattern differences vanish to a large extent by applying the declustering scheme but the resulting features again match with the degree in the network based on the corrected methods. This indicates that greater degree generally origins from the existence of long-distance connections and, thus, elevated average link distance.

Considering all of the above (node degree and average link distance), we emphasize the great impact of the application of the correction scheme and the choice of the link density. For the ECA, we have identified larger differences between the uncorrected and corrected versions particularly for the low link density, whereas the differences between the uncorrected and corrected ES have been substantial for elevated link density.

Local clustering coefficient

To further include an analysis of not only node-wise information of network topology (degree) and spatial distance but information on the neighborhood of a node, we examine the local clustering coefficient (Eq. 2.12). In this spirit, it can be considered as a higher-order network characteristic as compared with degree and average link distance.

In Fig. B.4, we show the local clustering coefficient patterns for a link density of $\rho = 2\%$. Comparing the pattern based on the different synchrony measures reveals notable differences. In the ES-based networks, there is a coherent region of low clustering coefficient in the southern Amazon basin which is not featured by the network based on the ECA. Additionally, Fig. B.4b,d illustrates that the declustering scheme does not lead to high similarity between the spatial network pattern in this

case. We assume that this is a direct consequence of the local clustering coefficient being a higher-order network measure. Whereas the declustering scheme corrects for node-wise biases, measuring the neighborhood of nodes and, thus, multi-node properties, reveals systematic differences between the event synchrony methods on which we base the networks.

As compared to the other two local network characteristics, the spatial pattern of the local clustering coefficient is notably finer even for the intermediate link density of $\rho = 2\%$. A further decrease of the link density (e.g., $\rho = 0.5\%$, not shown), results in a rather spiky and noisy pattern which does not allow for interpretation. Accordingly, the emergence of large-scale spatially coherent structures becomes increasingly unlikely with decreasing link density. This holds especially for regions with a low degree, where the existence of the few edges can lead to large variance in values of the local clustering coefficient. Therefore, we recommend an elevated link density for studying functional climate networks based on event synchrony utilizing higher-order network measures. Due to the large impact of the declustering scheme for higher link densities, this is particularly relevant when employing ES.

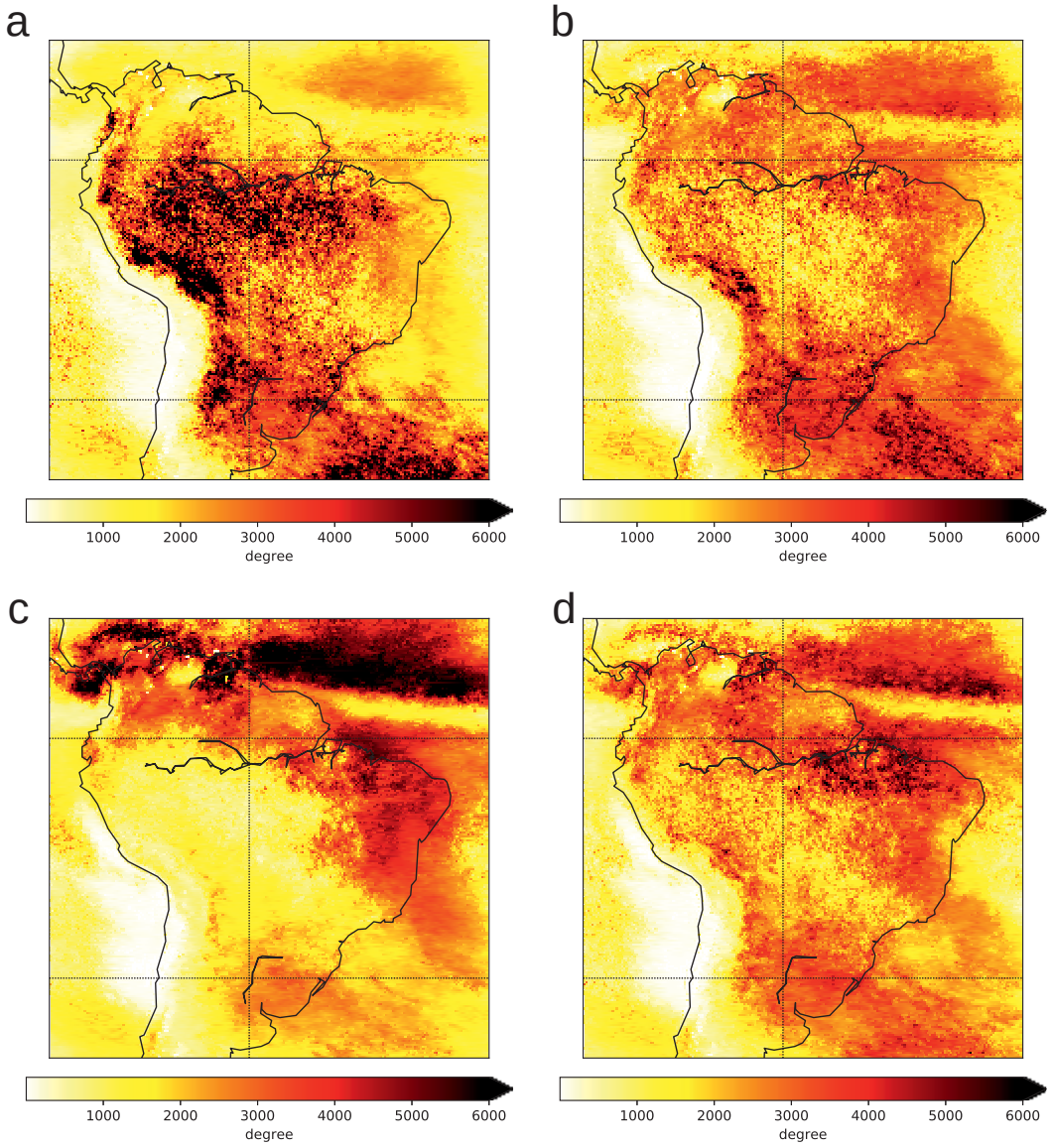

Figure B.2.: Same as Fig. B.1 but for a link density of $\rho = 5\%$

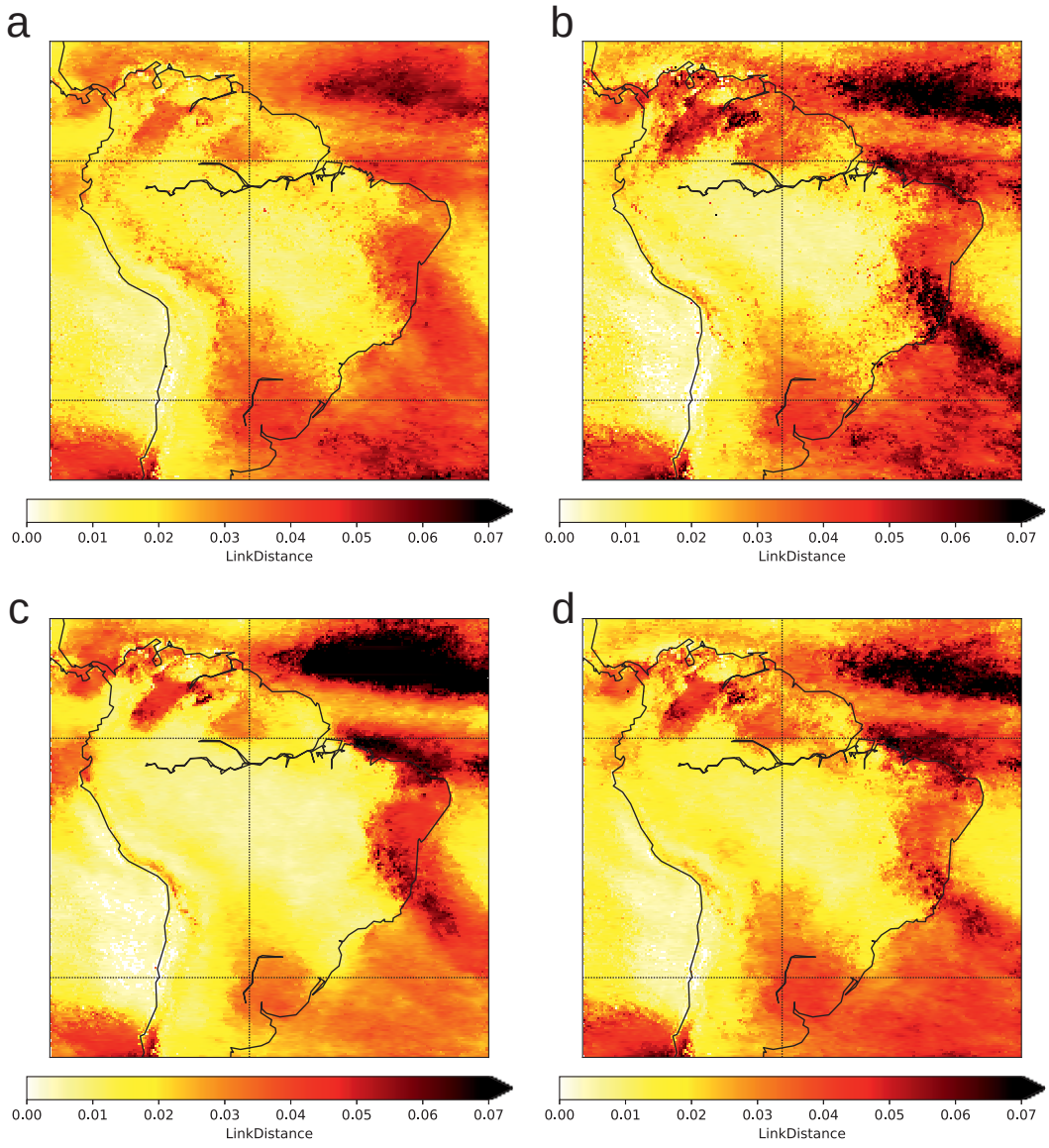

Figure B.3.: Same as Fig. B.1 but for the average link distance.

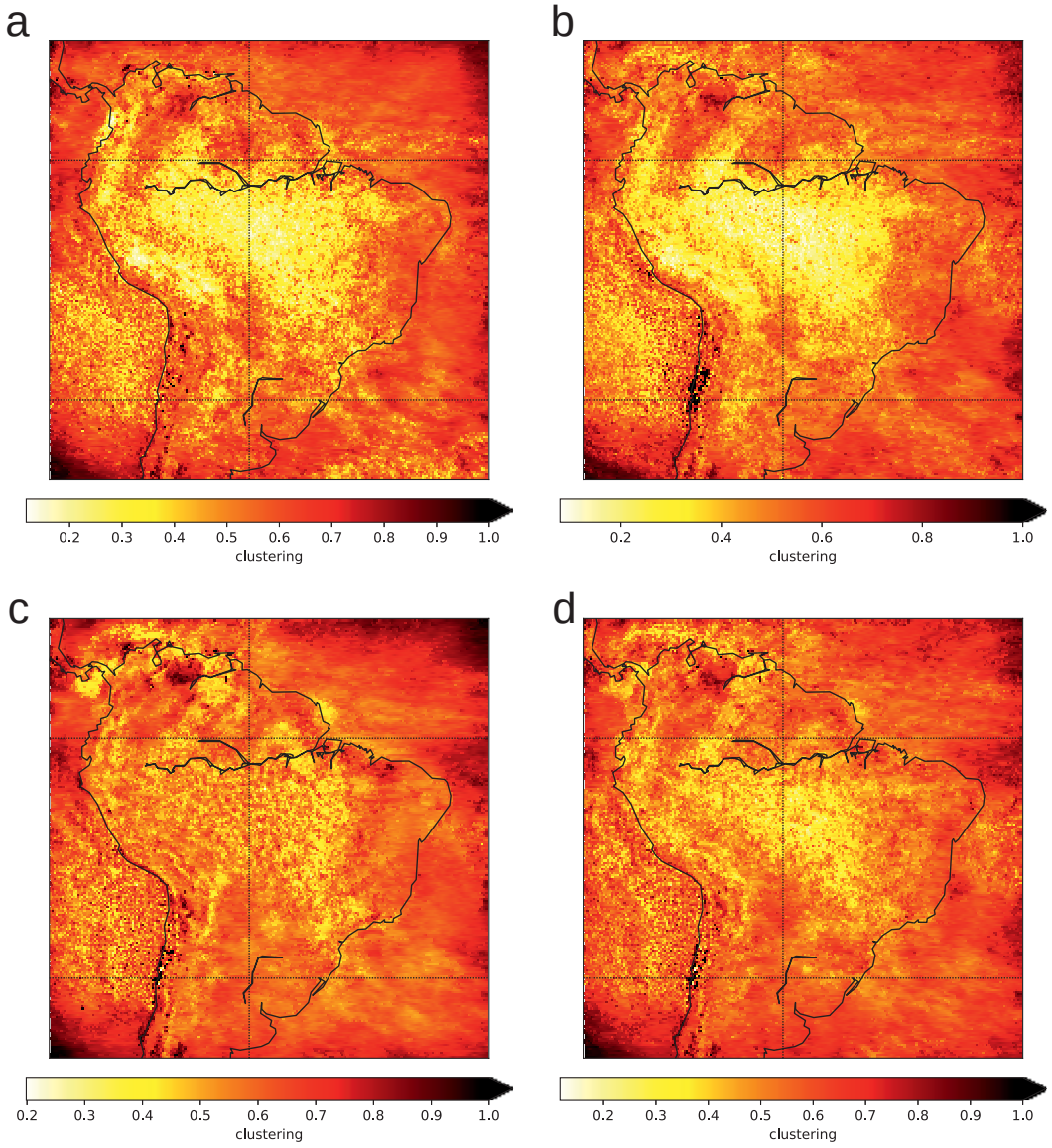

Figure B.4.: Same as Fig. B.1 but for the local clustering coefficient and a link density of $\rho = 2\%$

Single variable network measure correlations

In the following, we, as in chapter 4, address the interrelationships between the network measure values obtained utilizing the (un-)corrected versions of ES and ECA. As in Fig. 4.5, we show the values of the network measures based on the (un-)corrected versions of ES and ECA for each node colored by the corresponding pairing coefficient for two link densities in Figs. B.5 and B.6 as scatter plots. Thereby, we highlight the heterogeneity of the different effects resulting from the choice of the network measure (degree, average link distance, and local clustering coefficient), the link density, and the declustering scheme.

As described in the previous sections, there are minor qualitative differences in the network characteristics based on ECA and ES for low link density (see Fig. B.5). In line with the coherent trends towards higher similarity with incorporating the declustering scheme for degree and average link distance, we find a better quantitative agreement (i.e., an alignment around the line of identity in the scatter plot) in Fig. B.5b,f in comparison to Fig. B.5a,e. Although the scattered node degrees and average link distances show a clearer agreement after employing the declustering scheme, some differences remain. We suspect that these are sourced by the systematic differences between ECA and ES regarding the global/local coincidence interval. ECA counts all pairs of events that appear within a temporal distance of 3 days, whereas ES only counts those which fall into the dynamic coincidence interval, which can be smaller than 3 days. In our study setup, these minor differences do not allow for drawing different conclusions from the respective spatial pattern.

For an increased link density and the network measures based on the uncorrected event synchrony methods, we observe a broader scattering of the network measure values. In particular, we observe a clear separation for nodes with large and low pairing coefficient for degree and average link distance (Fig. B.6a,e). We explain this separation by the differently sourced biases for ES and ECA (ES: underestimation of nodes with high pairing due to event clusters, ECA: overestimation of nodes close to the ITCZ due to spatial clustering, see chapter 4). By incorporating the declustering scheme, the degree values get aligned along the diagonal (Fig. B.6b) which does not hold as clear for the average link distance (Fig. B.6f).

In line with the analysis of the spatial pattern of the local clustering coefficient, the distinction between values obtained by using ES and ECA is much less obvious than for the degree and average link distance and does not change markedly when the correction scheme is applied or the link density is varied. As a general tendency, we observe that nodes with a high pairing coefficient tend to exhibit higher local clustering values especially for larger link density in the ES-based network.

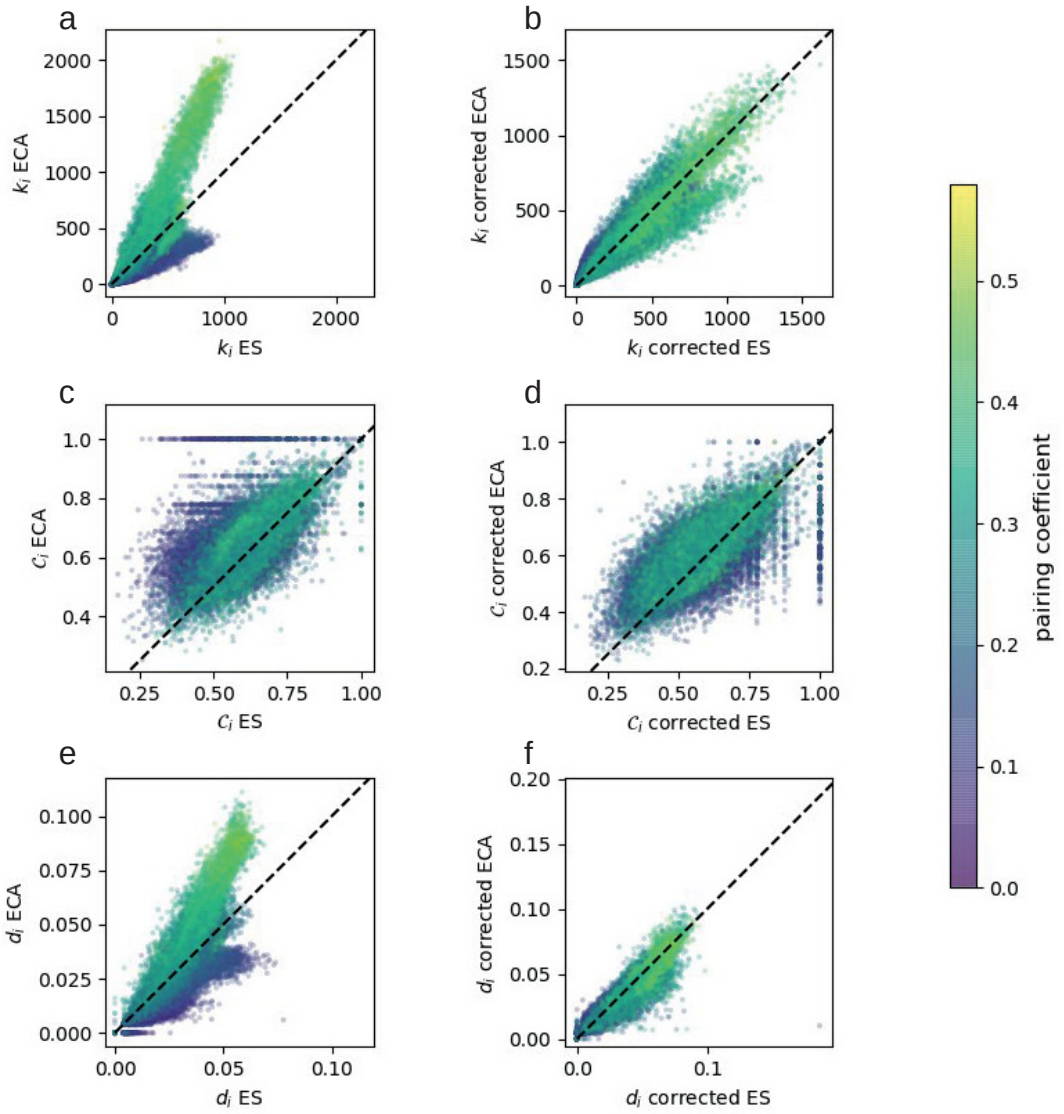

Figure B.5.: Values of (a,b) node degree, (c,d) local clustering coefficient and (e,f) average link distance for functional networks based on the two employed event synchrony measures. All results are shown for a link density of $\rho = 0.5\%$. Panels a,c,e (b,d,f) show the properties of the networks obtained without (with) incorporating the declustering scheme. Each individual value is color-coded by the pairing coefficient of the event series associated with the corresponding node. The lines of identity are indicated as black dashed lines. Note that unlike node degree and average link distance, the local clustering coefficient is restricted to the interval [0, 1], which leads to saturation effects in the plot.

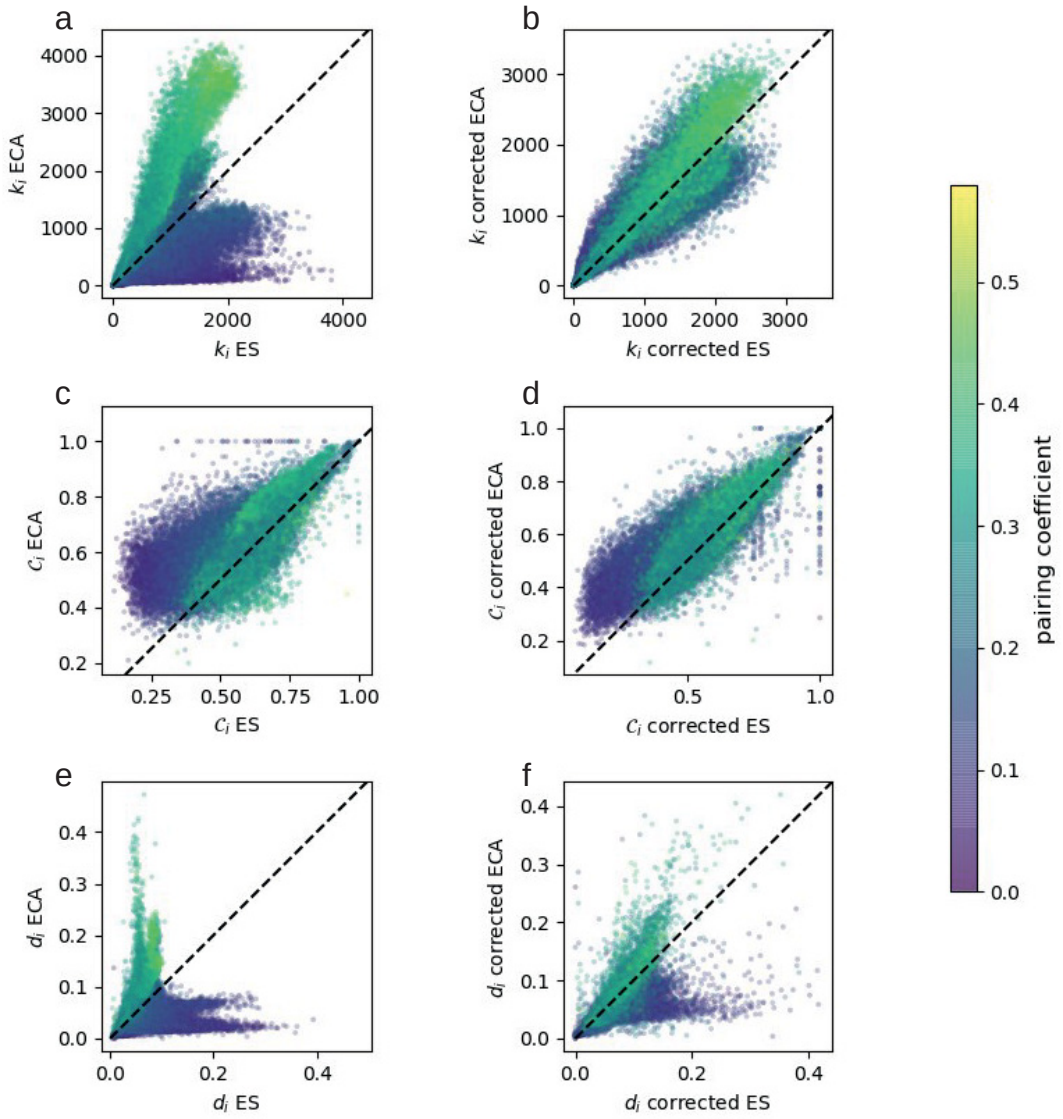

Figure B.6.: Same as Fig. B.5 but for a link density of $\rho = 2\%$.

Multi variable network measure correlations

Finally, we shed light on the interrelation between different network measures in Fig. B.7 and Fig. B.8.

Figure B.7 features the scatter plots for node degree and local clustering coefficient without (left) and with (right) incorporating the correction scheme. In all four panels, we observe a trend towards a negative correlation which has been observed also in previous studies [57, 59]. Besides this tendency, however, the overall scatter is fairly diffuse and consists of different groups of nodes sharing a certain dependency. Although we have no coherent interpretation of these structures, we assume that specific features (atmospheric, orographic, or geographic) cause differently shaped relations which result in the branches in Fig.B.7.

Figure B.8 illustrates the interrelation between node degree and average link distance. We find a general positive correlation of degree and link distance in all four setups which has been already indicated by the agreement in the respective spatial pattern. In the ES-based network, we however identify a gradual shift towards nodes with low pairing coefficient for high average link distances which is not shared by the ECA-based network. This feature is somewhat corrected by the declustering scheme but still partly remains for low degree and high average link distance.

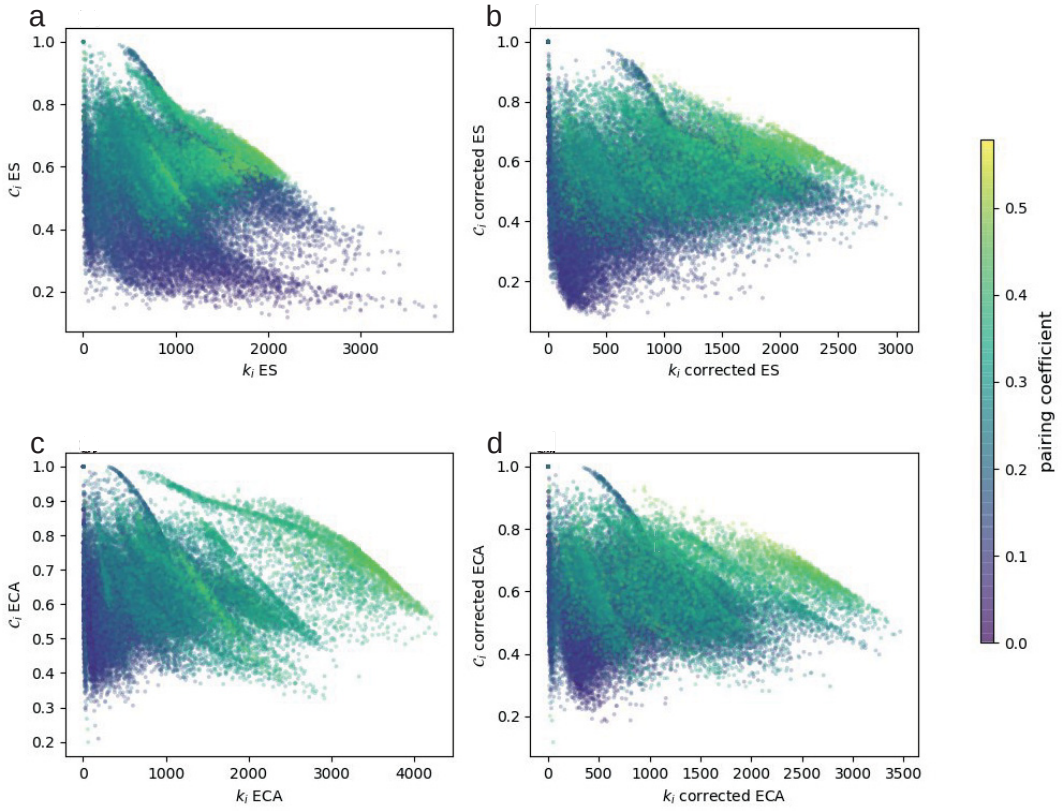

Figure B.7.: Relationship between node degree and local clustering coefficient for a link density of $\rho = 2\%$. Panels a,c (b,d) show the network measures without (with) incorporating the declustering scheme. The values are color coded according to the pairing coefficient of each node.

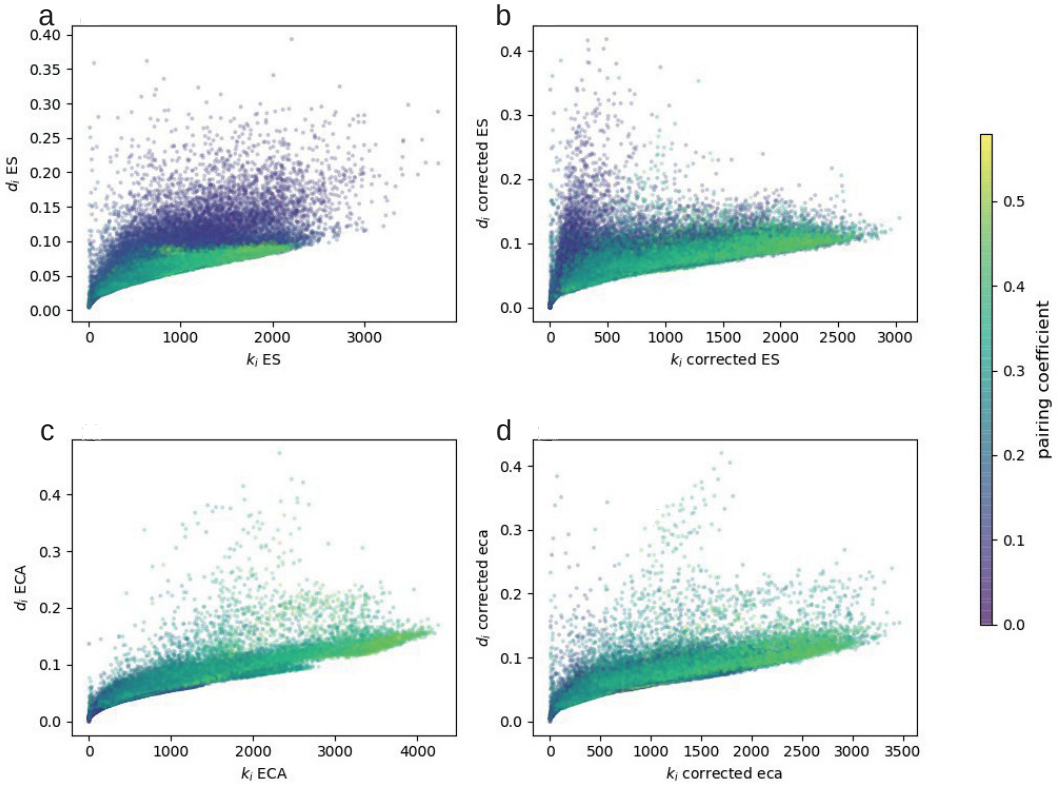

Figure B.8.: Same as Fig. B.7 but for degree and average link distance.

Summary

In this appendix, we have extended the case study 1 presented in chapter 4. We have not only considered three parameter settings of the link density ρ but also studied the degree along with the average link distance and the local clustering coefficient.

For the degree and average link distance, we have found that the differences between the spatial patterns disagree to greater extent when increasing the link density. By incorporating the declustering scheme, the patterns exhibit similar features. Additionally, we have identified a positive correlation of degree and average link distance in all analysis setups.

For the local clustering coefficient, which we have analyzed as an example for a higher-order network measure, the networks have exhibited distinct structures in networks based on ES and ECA. Also, the local clustering coefficient has appeared not to be as sensitive to the declustering scheme as the degree and the average link density. Whereas this might be related to systematic differences between the event synchrony measures (and related structural differences in the network topology), we have also observed that studying higher-order network measures typically requires larger link densities. As larger link densities tend to amplify (as mentioned before) the structural differences between ES and ECA, our results imply that selecting parameters and setting up the analysis framework has do be done by carefully considering the advantages and shortcomings of the respective event synchrony method, in particular when aiming for computing higher-order network measures.

Appendix C.

Additional information on ITCZ dynamics as seen by network theory - Jacknife and bootstrap test for clustering analysis

By splitting the data from 30 years into two parts cointaing 15 years, we produce two samples which we can individually analyze with our clustering framework. Following the steps we have conducted in chapter 6, we obtain two dendrograms (shown in Fig. C.1 and Fig. C.2) which we compare with the dendrogram computed based on the full time series (Fig. 6.3a).

Comparing Fig. C.1 with Fig. 6.3a, we find in general larger cophenetic distance between the models. This might be related to minor residues from the spin-up process. Although the dendrogram looks, therefore, slightly different, we can still observe main features of the full data analysis. First, the sub-cluster *MetUM-ENT, METUM-CTL, LMDZ5A, CAM4, ECHAM6.3* nearly completely matches the original cluster 1 (only ECHAM6.1 exhibit larger cophenetic distance and joins the CALTECH and AM2.1 model in the outlier group which we have also observed in the full time series analysis. Additionally, we find the cluster 2 models again in two sub-clusters. The only difference is the larger cophenetic distance between these two sub clusters.

Similarily to the above-stated analysis of the first 15 years, we can compare the clustering obtained from the second 15 years (shown in Fg. C.2) with the full data analysis. In contrast to the first Jacknife sample, we now identify CAM3 and CNRM-AM5 as two additional outliers (CALTECH and AM2.1 are again classified as single model clusters). Furthermore, we can again observe the sub-clusters which make up cluster 1 and cluster 2 of the full time series analysis. We do not obtain a similar grouping, as the cophenetic distance between these subgroups differs.

We conclude that the main features e.g. sub-clusters and outliers can be identified even using only half of the time series.

To further validate our results we have used a bootstrapping test: from the 30 years time series we randomly select 20 (shuffled) years which we treat as an individual time series. Repeating this for 20 realizations allows for computing the mean of the bootstrap data set. For each of the models, we subsequently compare the zonal mean network measures of the full data analysis with the mean of the bootstrapping test.

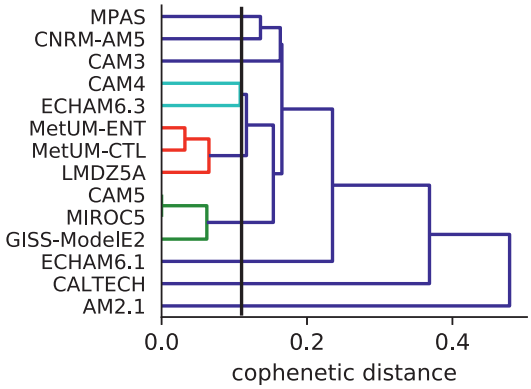

Figure C.1.: Dendrogram obtained from the global network analysis using the first 15 years of the time series. The vertical line indicates the level of cophenetic distance at which we have split the groups using the full time series.

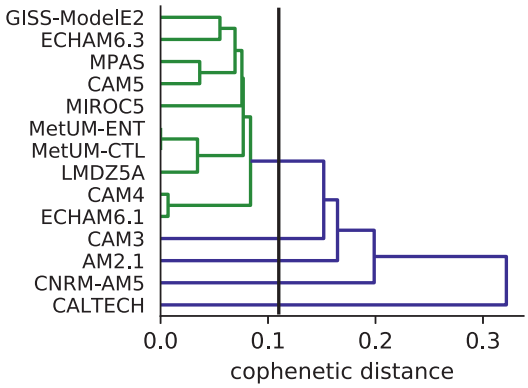

Figure C.2.: Same as Fig. C.1 but for using the second 15 years of the time series.

As this results in numerous figures (2 network measures $\times$ 14 models), we only show the result for the zonal mean degree of the LMDZ5A model as an example in Fig. C.3.

Figure C.3 shows that the mean features (double peak around ITCZ, additional maxima in the extratropics) are still present even if we randomly shuffle and restack the time series. We, therefore, consider that the zonal network measure distributions exhibit characteristic and distinct local maxima which certainly match the dynamics of this specific model.

Considering all of the above, we conclude that the results presented in chapter 6 are robust enough to draw the stated implications.

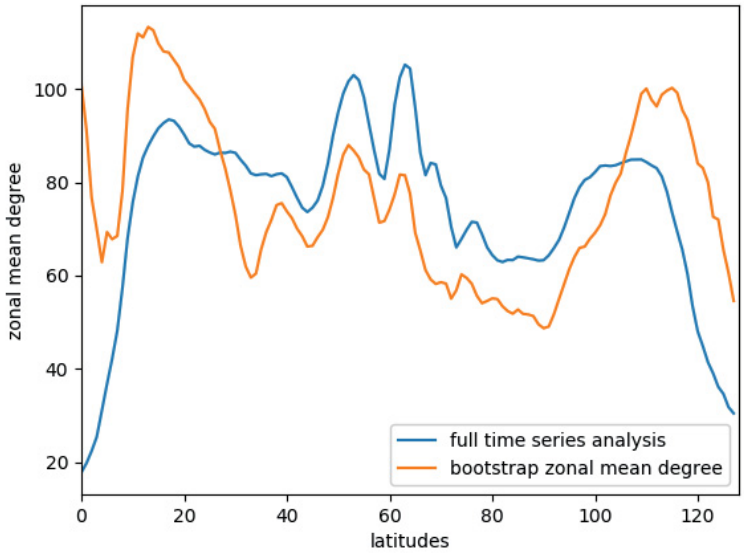

Figure C.3.: Bootstrap-mean zonal mean degree (orange) and zonal mean degree for the full data analysis (blue). Note that the y-axis features the 128 latitudes of the LMDZ5A model which correspond to the latitude range $[-90\circ, 90°]$.

Appendix D.

Robustness tests regarding the tweeting behavior changes presented in chapter 8

Raw inter-event times indicate differences between cohorts

The analysis presented in chapter 8 has been based on 200.000 users which have been active in April 2019. To verify our results, we have additionally crawled the tweeting behavior of 200.000 users which have been active in March 2019 and May 2019, respectively. In the following, we illustrate the raw data which we have utilized for all three samples. We do not show replicates of all figures based on the other data sets, but provide information on the inter-event times of both cohorts and clusters for all three data sets (shown in Fig. D.1). Note that we here also show the results for the cohort of 2012, which we have neglected in the main chapter. The distributions already indicate most of the results which have been presented in detail in chapter 8.

To highlight the differences in posting behavior on a sub-daily level for users starting at different points of time in the temporally changing environment of Twitter, we have analyzed the inter-event times of each user. Here, we again measure inter-event times within the tweeting timeseries of each user individually, before we compute the overall distribution for each cohort.

We show the distribution of inter-post times during the first year of each cohort in Fig. D.1 (left panels). The complex distributions have several features in common. All distributions exhibit a cut-off at inter post times of more than one year ($31536000s$) due to our analysis restriction, a global maximum at one day ($86400s$) followed by local maxima for multiple days and decreasing values for sub-daily inter-event times. The striking difference is the consecutive shift towards more sub-daily inter-event times for the cohorts which have entered twitter more lately and vice versa for super daily inter post times, whereas the global maxima of all distributions coincide. This trend is represented in all the data sets with the possible exception of extremely short inter-event times which have been observed with greater probability during the first year of cohort 4 from 2015 in sample 1 (top left panel in Fig. D.1). We suspect that these extreme very short inter-event times is most likely related to bots.

A larger fraction of short inter-event times in the first year automatically leads to a larger number of tweets per week for more recent cohorts (as shown in Fig.8.1). We, therefore, confirm an increasing activity of users which have entered Twitter more recently in all samples.

In chapter 8, we have split the users by their ratio of active days and obtained user types of distinct activity. Here, we repeat the analysis for sample 1 and sample 3 and show the inter-event time distribution of the user types in three samples in Fig. D.1 (right panels). Whereas the general features are the same as for the cohorts, the separation between the user types is larger (due to the criterion we used to obtain the user types). Nevertheless, the distributions directly relate to the boxplot in Fig. 8.2 which confirms distinct inter-event times for the user types.

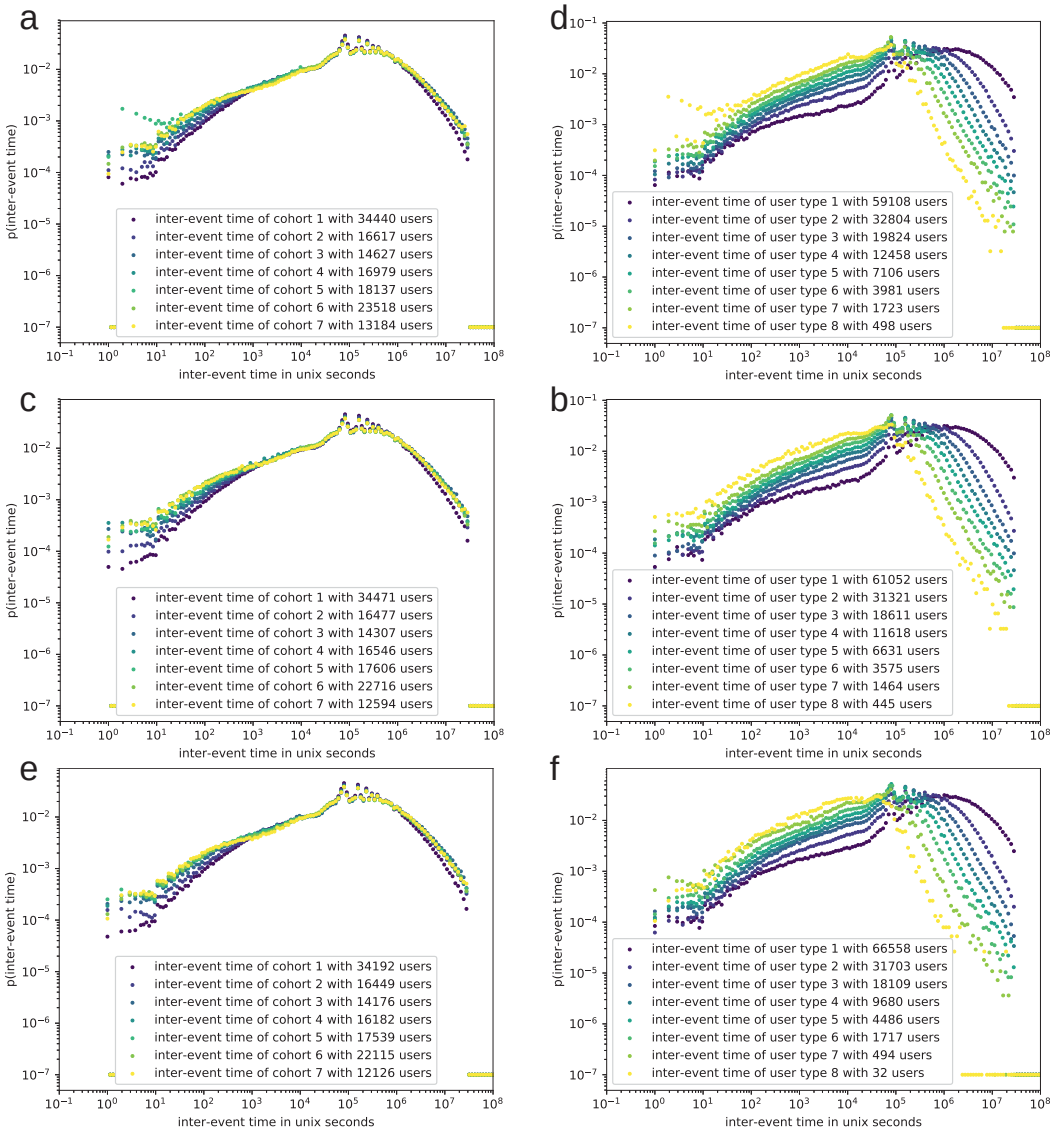

Figure D.1.: Inter-event time distribution for both cohorts (a,c,e) and user types (b,d,f) for the three analyzed samples (top to bottom). The distribution is colored differently for each cohort.

Cumulative starting day distributions and hashtag/URL use for the different user types

One striking finding of our analysis in chapter 8 has been that increasing fractions of recent cohorts end up as more active user types. While we have shown that for sample 2 (users active in April 2019) in Fig. 8.2, we here illustrate this in Fig. D.2a,b for the other samples. Figure D.2a,b depict the cumulative starting day distribution of users in the different user types for the study period. For both samples (sample 1, left; sample 3, right) the more active user types fill up in more recent time and, thus, exhibit a more recent median starting date. Furthermore, this indicates that more active user types grow quicker in more recent times.

As shown in Fig. D.1 and Fig. D.2, the trend of shortening inter-event times across cohorts and user types, as well as the increasing fraction of more recent users in more active user types is consistent for all samples. We, therefore, consider our results to be robust against the user selection.

To further confirm the trends in content use, we examine the average use of URLs and hashtags for the different user types (as studied in Fig. 8.3). The general trend (more hashtags and URLs per tweet in more recent years, see Fig. D.2c,d,e,f) is consistent with the described results in chapter 8.

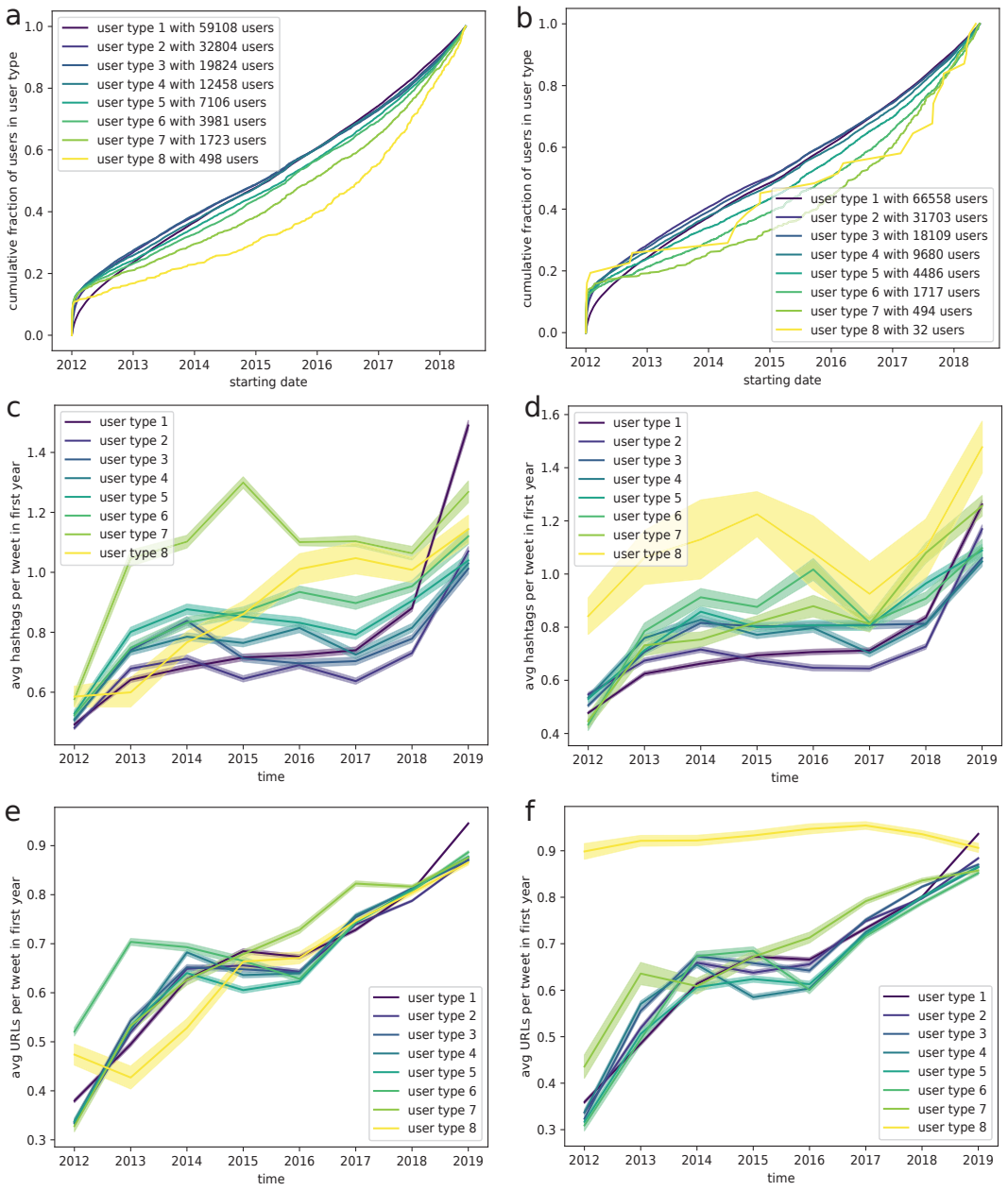

Figure D.2.: Cumulative starting day distribution of users in the user types for sample 1 (a) and sample 3 (b). Average hashtags per tweet over time per user type for sample 1 (c) and sample 3 (d). Average URLs per tweet over time per user type for sample 1 (e) and sample 3 (f). Standard deviation is indicated by shaded area.

Full 10% data set analysis

To this end, we want to stress that we do not find fundamentally different results using a data set featuring a denser sampling rate. As we have access to a 10% sample of tweeting behavior between 2012 and 2016, we also investigate the evolution of tweeting behavior of a smaller subsample of users for this shorter period. We show the inter-event time for the according four cohorts (2013-2016) in Fig. D.3. Although the trend of increasing short inter-event times is visible to lesser extent due to the brief study period, our findings regarding the increase of cohort activity are also robust in this temporal domain employing a higher sampling rate.

In addition to studying the cohorts, we have, again, split the users into distinct user types by means of the ratio of active days. Thereafter, we have determined the composition of the user types by computing the fractions of the four cohorts in the respective user types. As in chapter 8, Fig. D.4a shows the fractions of the four cohorts in four histograms. We identify an increasing fraction of more recent users in the more active user types. Furthermore, the user types separate well regarding their quantiles in the inter-event time distributions Fig. D.4b. All these observations agree with the results obtained employing the full period 1% sample.

Due to the overall similar results of the 1% and the 10% sample, we conclude that the resolution of 1% is sufficient to draw the conclusions we have presented in chapter 8.

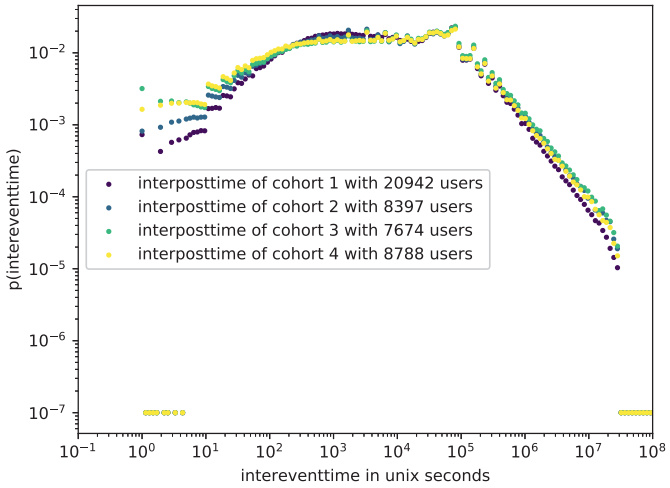

Figure D.3.: Inter-event time distribution for each cohort of the 10% user sample.

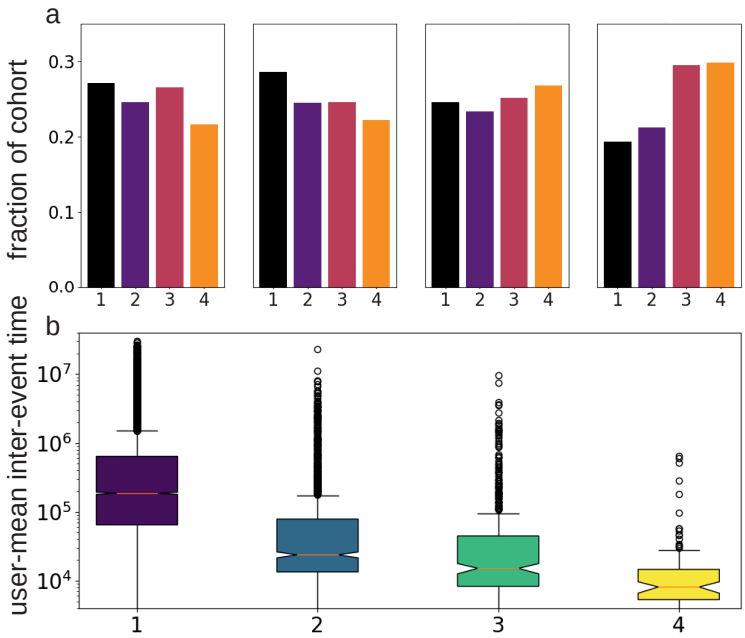

Figure D.4.: Fractions of cohorts for each user type (a). Each histogram sums up to 1. User-mean inter-event time distribution for the user types (b).

Code and data availability

The code for the data preprocessing, data analysis and data visualization in this thesis is available from the author upon request. Please do not hesitate to contact me. All data for the work in part 1 and part 2 is available online and was referenced accordingly. The data used in part 3 was provided by external sources and can be provided by the author upon request.

Bibliography

[1] M. Hilbert and P. Lopez. The world's technological capacity to store, communicate, and compute information. *Science*, 332, (2011), 60–66.

[2] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Phys. Rev. E*, 51 (5), (1995), 4282–4286.

[3] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.*, 81 (2), (2009), 501–638.

[4] V. Dhar. Data Science and Prediction. *Commun. ACM*, 56 (12), (2013), 64–73.

[5] N. Boers, B. Bookhagen, H. M. J. Barbosa, N. Marwan, J. Kurths, and J. A. Marengo. Prediction of extreme floods in the eastern Central Andes based on a complex networks approach. *Nat. Commun.*, 5, (2014), 5199.

[6] D. Sornette. *Why stock markets crash: critical events in complex financial systems*. Princeton University Press, 2017.

[7] D. Dougherty and D. D. Dunne. Digital Science and Knowledge Boundaries in Complex Innovation. *Organ. Sci.*, 23 (5), (2012), 1467–1484.

[8] N. Eagle and A. S. Pentland. Reality mining : sensing complex social systems. *Pers. Ubiquitous Comput.*, 10 (4), (2006), 255–268.

[9] F. Provost and T. Fawcett. Data science and its relationship to big data and data-driven decision making. *Data Sci. big data*, 1 (1), (2013).

[10] A.-L. Barabási and M. Posfai. *Network Science*. Cambridge University Press, 2016.

[11] L. C. Freeman. Centrality in Social Networks -Conceptual Clarification. *Soc. Networks*, 1, (1978), 215–239.

[12] N. Mohajeri, J. R. French, and A. Gudmundsson. Entropy measures of street-network dispersion: Analysis of coastal cities in Brazil and Britain. *Entropy*, 15 (9), (2013), 3340–3360.

[13] J. Maluck and R. V. Donner. A Network of Networks Perspective on Global Trade. *PLoS One*, 10 (7), (2015), e0133310.

[14] M. G. Rosenblum, A. S. Pikovsky, and J. Kurths. Synchronization approach to analysis of biological systems. *Fluct. Noise Lett.*, 4 (1), (2004), 53–62.

[15] S. H. Y. Chan, R. V. Donner, and S. Laemmer. Urban road networks – spatial networks with universal geometric features? *Eur. Phys. J. B*, 84 (4), (2011), 563–577.

[16] H. A. Dijkstra. *Nonlinear Climate Dynamics*. Cambridge University Press, 2013.

[17] V. Lucarini, R. Blender, C. Herbert, F. Ragone, S. Pascale, and J. Wouters. Mathematical and physical ideas for climate science. *Rev. Geophys.*, 52, (2014), 809–859.

[18] R. A. McLeman. *Climate and human migration: Past experiences, future challenges*. Cambridge University Press, 2014.

[19] M. McCormick et al. Climate change during and after the Roman Empire: Reconstructing the past from scientific and historical evidence. *J. Interdiscip. Hist.*, 43 (2), (2012), 169–220.

[20] L. M. Hunter, J. K. Luna, and R. M. Norton. Environmental Dimensions of Migration. *Annu. Rev. Sociol.*, 41, (2015), 377–397.

[21] R. W. Katz and B. B. Brown. Extreme Events in a changing climate, variability is more important than averages. *Clim. Chang.*, 21, (1992), 289–302.

[22] D. R. Easterling, J. L. Evans, P. Y. Groisman, T. R. Karl, K. E. Kunkel, and P Ambenje. Observed Variability and Trends in Extreme Climate Events : A Brief Review. *Bull. Am. Meteorol. Soc.*, 81 (3), (1999), 417–426.

[23] S. Bony et al. Clouds , circulation and climate sensitivity. *Nat. Publ. Gr.*, 8, (2015), 261–268.

[24] N. Boers, B. Goswami, A. Rheinwalt, B. Bookhagen, B. Hoskins, and J. Kurths. Complex networks reveal global pattern of extreme-rainfall teleconnections. *Nature*, 566, (2019), 373–377.

[25] DISC GES. TRMM. *TRMM Rainfall Estim. L3 3 hour 0.25 degree x 0.25 degree V7*, 25, (2016).

[26] A. A. Tsonis and P. J. Roebber. The architecture of the climate network. *Phys. A*, 333, (2004), 497–504.

[27] A. A. Tsonis, K. L. Swanson, and P. J. Roebber. What do networks have to do with climate? *Bull. Am. Meteorol. Soc.*, 87 (5), (2006), 585–595.

[28] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. The backbone of the climate network. *EPL*, 87 (4), (2009), 48007.

[29] N. Boers, B. Bookhagen, N. Marwan, J. Kurths, and J. A. Marengo. Complex networks identify spatial patterns of extreme rainfall events of the South American Monsoon System. *Geophys. Res. Lett.*, 40, (2013), 4386–4392.

[30] R. V. Donner, M. Wiedermann, and J. F. Donges. *Complex Network Techniques for Climatological Data Analysis*. Ed. by C Franzke and T O'Kane. 1st ed. Cambridge: Cambridge University Press, 2017, pp. 159–183.

[31] M. Wiedermann, A. Radebach, J. F. Donges, J. Kurths, and R. V. Donner. A climate network-based index to discriminate different types of El Niño and La Niña. *Geophys. Res. Lett.*, 43 (13), (2016), 7176–7185.

[32] V. Stolbova, E. Surovyatkina, B. Bookhagen, and J. Kurths. Tipping elements of the Indian monsoon : Prediction of onset and withdrawal. *Geophys. Res. Lett.*, 43, (2016), 3982–3990.

[33] R. Q. Quiroga, T. Kreuz, and P. Grassberger. Event synchronization: A simple and fast method to measure synchronicity and time delay patterns. *Phys. Rev. E*, 66 (4), (2002), 041904.

[34] N. Malik, B. Bookhagen, N. Marwan, and J. Kurths. Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks. *Clim. Dyn.*, 39, (2012), 971–987.

[35] Y Liu, C Kliman-Silver, and A Mislove. The Tweets They Are a-Changin: Evolution of Twitter Users and Behavior. *ICWSM*, 30 (314), (2014).

[36] R. De. Societal impacts of information and communications technology. *IIMB Manag. Rev.*, 28 (2), (2016), 111–118.

[37] D. Lazer et al. Computational Social Science. *Science*, 323, (2009), 721–724.

[38] J. Borge-Holthoefer et al. Structural and Dynamical Patterns on Online Social Networks : The Spanish May 15th Movement as a Case Study. *PLoS One*, 6 (8), (2011), e23883.

[39] V. Sekara, A. Stopczynski, and S. Lehmann. Fundamental structures of dynamic social networks. *PNAS*, 113 (36), (2016), 9977–9982 SEE.

[40] M. Rosvall and C. T. Bergstrom. Mapping Change in Large Networks. *PLoS One*, 5 (1), (2010), e8694.

[41] P. Holme and J. Saramäki. Temporal networks. *arXiv:1108.1780v1*, (2011). arXiv: `arXiv:1108.1780v2`.

[42] H. Rosa. *Social acceleration: A new theory of modernity*. Columbia University Press, 2013.

[43] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359, (2018), 1146–1151.

[44] A. J. Stewart, M. Mosleh, M. Diakonova, A. A. Arechar, D. G. Rand, and J. B. Plotkin. Information gerrymandering and undemocratic decisions. *Nature*, 573, (2019), 117–121.

[45] S. Achard, R. Salvador, B. Whitcher, J. Suckling, and E. T. Bullmore. A Resilient, Low-Frequency, Small-World Human Brain Functional Network with Highly Connected Association Cortical Hubs. *J. Neurosci.*, 26 (1), (2006), 63–72.

[46] P. E. Vertes, A. F. Alexander-bloch, N. Gogtay, J. N. Giedd, J. L. Rapoport, and E. T. Bullmore. Simple models of human brain functional networks. *PNAS*, 109 (15), (2012), 5868–5873.

[47] I. Joncour, E. Moraux, G. Duchene, and L. Murdy. Multiscale Spatial Analysis of Young Stars Complex using the dbscan Clustering Algorithm. *Proc. Astron. Data Anal. Softw. Syst. XXVIII. ASP Conf. Ser.*, 523, (2019), 87.

[48] F. Herten and G. Waldmann. Functional principles of early time measurement at Stonehenge and Nebra. *Archäologische Informationen*, 41, (2018), 275–287.

[49] A Einstein. Uber einen die Erzeugung und Verwandlung des Lichtes betreffenden heurischen Gesichtpunkt. *Ann. Phys.*, 17, (1905), 132–148.

[50] L Euler. Solutio Problemat is ad Geometriam Situs Pertinentis. *Comment. Acad. Sci. Imp. Petropolitanae*, 8, (1741), 128–140.

[51] M. Bastian, S. Heymann, and M. Jacomy. Gephi : An Open Source Software for Exploring and Manipulating Networks. *Third Int. ICWSM Conf.*, (2009).

[52] S. H. Strogatz. Exploring complex networks. *Nature*, 410 (6825), (2001), 268.

[53] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang. Complex networks: Structure and dynamics. *Phys. Rep.*, 424 (4-5), (2006), 175–308.

[54] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74 (1), (2002), 47.

[55] N. Boers, R. V. Donner, and B. Bookhagen. Complex network analysis helps to identify impacts of the El Niño Southern Oscillation on moisture divergence in South America. *Clim. Dyn.*, 45, (2014), 619–632.

[56] J. F. Donges, H. C. Schultz, N. Marwan, Y. Zou, and J. Kurths. Investigating the topology of interacting networks: Theory and application to coupled climate subnetworks. *Eur. Phys. J. B*, 84 (4), (2011), 635–651. arXiv: 1102.3067.

[57] A. Radebach, R. V. Donner, J. Runge, J. F. Donges, and J. Kurths. Disentangling different types of El Niño episodes by evolving climate network analysis. *Phys. Rev. E*, 88 (5), (2013), 052807.

[58] R. Cazabet, H. Takeda, M. Hamasaki, and F. Amblard. Using dynamic community detection to identify trends in user-generated content. *Soc. Netw. Anal. Min.*, 2, (2012), 361–371.

[59] M. Wiedermann, J. F. Donges, D. Handorf, J. Kurths, and R. V. Donner. Hierarchical structures in Northern Hemispheric extratropical winter ocean–atmosphere interactions. *Int. J. Climatol.*, 37 (10), (2017), 3821–3836.

[60] S. Asur, S. Parthasarathy, and D. Ucar. An Event-based Framework for Characterizing the Evolutionary Behavior of Interaction Graphs. *ACM Trans. Comput. Log.*, 2 (3), (2001), 1–35.

[61] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435, (2005), 814–818.

[62] T. Aynaud, J.-l. Guillaume, Q. Wang, and E. Fleury. *Communities in Evolving Networks: Definitions, Detection, and Analysis Techniques*. 2nd ed. Springer, 2013, pp. 159–200.

[63] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, P10008, (2008).

[64] M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci.*, 103 (23), (2006), 8577–8582. arXiv: `0602124 [physics]`.

[65] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *PNAS*, 105 (4), (2008), 1118–1123.

[66] N. Djurdjevac, S. Bruckner, T. O. F. Conrad, and C. Schuette. Random Walks on Complex Modular Networks 12. *JNAIAM*, 6 (1-2), (2011), 29–50.

[67] M. Sarich, N. Djurdjevac, S. Bruckner, T. O. F. Conrad, and C. Schuette. Modularity revisited: a novel dynamics-based concept for decomposing complex networks. *J. Comput. Dyn.*, 1 (1), (2014), 191–212.

[68] T. P. Peixoto. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. *Phys. Rev. X*, 4 (1), (2014), 011047.

[69] T. P. Peixoto. Reconstructing Networks with Unknown and Heterogeneous Errors. *Phys. Rev. X*, 8 (4), (2018), 41011.

[70] V. Stolbova, P. Martin, B. Bookhagen, N. Marwan, and J. Kurths. Topology and seasonal evolution of the network of extreme precipitation over the Indian subcontinent and Sri Lanka. *Nonlinear Process. Geophys.*, 21, (2014), 901–917.

[71] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe. A framework for community identification in dynamic social networks. *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, (2007), 717–726.

[72] D. Brockmann and D. Helbing. The Hidden Geometry of Complex , Network-Driven Contagion Phenomena. *Science*, 342, (2013), 1337–1343.

[73] J. F. Donges, C. F. Schleussner, J. F. Siegmund, and R. V. Donner. Event coincidence analysis for quantifying statistical interrelationships between event time series. *Eur. Phys. J. Spec. Top.*, 225, (2016), 471–487.

[74] J. Benesty, J. Chen, Y. Huang, and I. Cohen. Pearson correlation coefficient. In: *Noise reducting in speech processing*. 2009, pp. 1–4.

[75] U. Ozturk, N. Malik, K. Cheung, N. Marwan, and J. Kurths. A network - based comparative study of extreme tropical and frontal storm rainfall over Japan. *Clim. Dyn.*, 53 (1-2), (2019), 521–532.

[76] A. Agarwal, N. Marwan, M. Rathinasamy, B. Merz, and J. Kurths. Multi-scale event synchronization analysis for unravelling climate processes: a wavelet-based approach. *Nonlinear Process. Geophys.*, 24, (2017), 599–611.

[77] A. Odenweller and R. V. Donner. Disentangling synchrony from serial dependency in paired-event time series. *Phys. Rev. E*, 101 (5), (2020), 052213.

[78] F. Hassanibesheli and R. V. Donner. Network inference from the timing of events in coupled dynamical systems. *Chaos*, 29 (083125), (2019).

[79] M. Barthelemy. Spatial Networks. *Phys. Rep.*, 499 (1-3), (2011), 1–101.

[80] N. Molkenthin, H. Kutza, L. Tupikina, N. Marwan, J. F. Donges, U. Feudel, and R. V. Donner. Edge anisotropy and the geometric perspective on flow networks. *Chaos*, 27 (3), (2017), 035802.

[81] J. F. Donges, Y Zou, N Marwan, and J Kurths. Complex networks in climate dynamics. *Eur. Phys. J. Spec. Top.*, 174, (2009), 157–179.

[82] J Heitzig, J. F. Donges, Y Zou, N Marwan, and J Kurths. Node-weighted measures for complex networks with spatially embedded, sampled, or differently sized nodes. *Eur. Phys. J. B*, 85 (1), (2012), 38.

[83] M. Wiedermann, J. F. Donges, J. Heitzig, and J Kurths. Node-weighted interacting network measures improve the representation of real-world complex systems. *EPL*, 102 (2), (2013), 28007.

[84] D. C. Zemp, M. Wiedermann, J. Kurths, A. Rammig, and J. F. Donges. Node-weighted measures for complex networks with directed and weighted edges for studying continental moisture recycling. *EPL*, 107 (5), (2014), 58005.

[85] A. Rheinwalt, N. Marwan, J Kurths, P. Werner, and F.-W. Gerstengabe. Boundary effects in network measures of spatially embedded networks. *EPL*, 100 (2), (2012), 28002.

[86] A. Voigt et al. The tropical rain belts with an annual cycle and a continent model intercomparison project: TRACMIP. *J. Adv. Model. Earth Syst.*, 8 (4), (2016), 1868–1891.

[87] B. Stevens et al. Atmospheric component of the MPI-M earth system model: ECHAM6. *J. Adv. Model. Earth Syst.*, 5 (2), (2013), 146–172.

[88] D. P. Dee et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q.J.R. Meteorol. Soc*, 137, (2011), 553–597.

[89] N. S. Keenlyside and M. Latif. Understanding equatorial atlantic interannual variability. *J. Clim.*, 20 (1), (2007), 131–142.

[90] M. Á. Serrano and M. Boguñá. Topology of the world trade web. *Phys. Rev. E*, 68 (1), (2003), 015101.

[91] M. Lenzen, K. Kanemoto, D. Moran, and A. Geschke. Mapping the structure of the world economy. *Environ. Sci. Technol.*, 46 (15), (2012), 8374–8381.

[92] A. Barrat, M. Barthelemy, and A. Vespignani. Weighted Evolving Networks: Coupling Topology and Weight Dynamics. *Phys. Rev. Lett.*, 92 (22), (2004), 228701.

[93] M. T. Gastner and M. E. J. Newman. The spatial structure of networks. *Eur. Phys. J. B*, 49 (2), (2006), 247–252.

[94] C. Rosenzweig, A. Iglesias, X. B. Yang, P. R. Epstein, and E. Chivian. Climate change and extreme weather events. *Glob. Chang. Hum. Heal.*, 2 (2), (2001), 90–104.

[95] N. Malik, N. Marwan, and J. Kurths. Nonlinear Processes in Geophysics Spatial structures and directionalities in Monsoonal precipitation over South Asia. *Nonlinear Process. Geophys.*, 17, (2010), 371–381.

[96] N. Boers, B. Bookhagen, N. Marwan, and J. Kurths. Spatiotemporal characteristics and synchronization of extreme rainfall in South America with focus on the Andes Mountain range. *Clim. Dyn.*, 46, (2015), 601–617.

[97] N. Boers, A. Rheinwalt, B. Bookhagen, H. M. J. Barbosa, N. Marwan, J. A. Marengo, and J. Kurths. The South American rainfall dipole: A complex network analysis of extreme events. *Geophys. Res. Lett.*, 41 (20), (2014), 7397–7405.

[98] N. Boers, B. Bookhagen, J. A. Marengo, N. Marwan, J.-S. von Storch, and J. Kurths. Extreme Rainfall of the South American Monsoon System : A Dataset Comparison Using Complex Networks. *J. Clim.*, 28, (2015), 1031–1056.

[99] J. Zhou and K. M. Lau. Does a Monsoon Climate Exist over South America ? *J. Clim.*, 11, (1998), 1020–1040.

[100] B. Bookhagen and M. R. Strecker. Orographic barriers , high-resolution TRMM rainfall , and relief variations along the eastern Andes. *Geophys. Res. Lett.*, 35 (Lo6403), (2008), 1–6.

[101] L. M. V. Carvalho, C. Jones, and B Liebmann. Extreme Precipitation Events in Southeastern South America and Large-Scale Convective Patterns in the South Atlantic Convergence Zone. *J. Clim.*, 15, (2002), 2377–2394.

[102] J. A. Marengo and W. R. Soares. Climatology of the Low-Level Jet East of the Andes as Derived from the NCEP – NCAR Reanalyses : Characteristics and Temporal Variability. *J. Clim.*, 17 (12), (2004), 2261–2280.

[103] B Liebmann, G. N. Kiladis, C. S. Vera, A. C. Saulo, and L. M. V. Carvalho. Subseasonal Variations of Rainfall in South America in the Vicinity of the Low-Level Jet East of the Andes and Comparison to Those in the South Atlantic Convergence Zone. *J. Clim.*, 17, (2004), 3829–3842.

[104] P. Salio, M. Nicolini, and E. J. Zipser. Mesoscale Convective Systems over Southeastern South America and Their Relationship with the South American Low-Level Jet. *Mon. Weather Rev.*, 135, (2007), 1290–1309.

[105] J. D. Durkee, T. L. Mote, and J. M. Shepherd. The Contribution of Mesoscale Convective Complexes to Rainfall across Subtropical South America. *J. Clim.*, 22, (2009), 4590–4605.

[106] M. Gelbrecht, N. Boers, and J. Kurths. Phase coherence between precipitation in South America and Rossby waves. *Sci. Adv.*, 8 (12), (2018).

[107] C. Vera et al. Toward a Unified View of the American Monsoon Systems. *J. Clim. - Spec. Sect.*, 19, (2006), 4977–5000.

[108] K. Ninomiya and H. Muraki. Large-Scale Circulations over East Asia during Baiu Period of 1979. *J. Meteorlogical Soc. Japan*, 64 (3), (1986), 409–429.

[109] R. Krishnan and M. Sugi. Baiu Rainfall Variability and Associated Monsoon Teleconnections. *J. Meteorlogical Soc. Japan*, 79 (3), (2001), 851–860.

[110] H. Ueda and T. Yasunari. Abrupt Seasonal Change of Large-Scale Convective Activity. *J. Meteorlogical Soc. Japan*, 73 (4), (1995), 795–809.

[111] Y. Okada and K. Yamazaki. Climatological Evolution of the Okinawa Baiu and Differences in Large-Scale Features during May and June. *J. Clim.*, 25, (2012), 6287–6303.

[112] P. X. Wang, B. Wang, H. Cheng, J. Fasullo, Z. T. Guo, T. Kiefer, and Z. Y. Liu. The global monsoon across time scales: Mechanisms and outstanding issues. *Earth-Science Rev.*, 174, (2017), 84–121.

[113] D. Yihui and J. C. L. Chan. The East Asian summer monsoon : an overview. *Meteorol. Atmos. Phys.*, 89, (2005), 117–142.

[114] E. Fukui. Distribution of extraordinarily heavy rainfalls in Japan. *Geogr. Rev. Jpn.*, 43 (10), (1970), 581–593.

[115] B. Preethi, M. Mujumdar, R. H. Kripalani, A. Prabhu, and R. Krishnan. Recent trends and tele - connections among South and East Asian summer monsoons in a warming environment. *Clim. Dyn.*, 48 (7), (2017), 2489–2505.

[116] H. Li, S. He, K. Fan, and H. Wang. Relationship between the onset date of the Meiyu and the South Asian anticyclone in April and the related mechanisms. *Clim. Dyn.*, 52 (1-2), (2018), 209–226.

[117] K. Suda and T. Asakura. A study on the Unusual Baiu Season in 1954 by Means of Northern Atmosphere Upper Air Mean Charts. *J. Meteorlogical Soc. Japan*, 33 (6), (1955), 233–244.

[118] T. Sampe and S.-P. Xie. Large-Scale Dynamics of the Meiyu-Baiu Rainband : Environmental Forcing by the Westerly Jet. *J. Clim.*, 23 (1), (2010), 113–134.

[119] X Zhu, Z Wu, and J He. Anomalous Meiyu onset averaged over the Yangtze River valley. *Theor. Appl. Climatol.*, 94, (2008), 81–95.

[120] K.-S. Choi, B. Wang, and D.-W. Kim. Changma onset definition in Korea using the available water resources index and its relation to the Antarctic oscillation. *Clim. Dyn.*, 38, (2012), 547–562.

[121] J. Zhu, D.-Q. Huang, Y.-C. Zhang, A.-N. Huang, X.-Y. Kuang, and Y. Huang. Decadal changes of Meiyu rainfall around 1991 and its relationship with two types of ENSO. *J. Geophys. Res. Atmos.*, 118, (2013), 9766–9777.

[122] E Kalnay et al. The NCEP / NCAR 40-Year Reanalysis Project. *Bull. Am. Meteorol. Soc.*, 77 (3), (1996), 437–472.

[123] B. Y. M. Kanamitsu, W. Ebisuzaki, W. Jack, S.-K. Yang, J. J. Hnilo, M Fiorino, and G. L. Potter. NCEP-DOE AMIP-II Reanalysis (R-2). *Bull. Am. Meteorol. Soc.*, 83 (11), (2002), 1631–1644.

[124] G. T.-J. Chen. Large-Scale Circulations Associated with the East Summer Monsoon and the Mei-Yu over South China and Taiwan. *J. Meteorlogical Soc. Japan*, 72 (6), (1994), 959–983.

[125] Y. Liu and Y. Ding. Teleconnection between the Indian summer monsoon onset and the Meiyu over the Yangtze River Valley. *Sci. China Ser. D Earth Sci.*, 51 (7), (2008), 1021–1035.

[126] T. Tomita, T. Yamaura, and T. Hashimoto. Interannual Variability of the Baiu Season near Japan Evaluated from the Equivalent Potential Temperature. *J. Meteorlogical Soc. Japan*, 89 (5), (2011), 517–537.

[127] P. Guan, G. Chen, W. Zeng, and Q. Liu. Corridors of Mei-Yu-Season Rainfall over Eastern China. *J. Clim.*, 23, (2020), 2603–2626.

[128] J. Tan, C. Jakob, W. B. Rossow, and G. Tselioudis. Increases in tropical rainfall driven by changes in frequency of organized deep convection. *Nature*, 519, (2015), 451–454.

[129] J. Wallace and P. Hobbs. *Atmospheric Science*. Academic Press, 2006.

[130] S. M. Kang, Y. Shin, and S.-P. Xie. Extratropical forcing and tropical rainfall distribution : energetics framework and ocean Ekman advection. *NPJ Clim. Atmos. Sci.*, 1 (20172), (2018), 1–10.

[131] S. M. Kang, D. M. W. Frierson, and I. M. Held. The Tropical Response to Extratropical Thermal Forcing in an Idealized GCM : The Importance of Radiative Feedbacks and Convective Parameterization. *J. Atmos. Sci.*, 66, (2009), 2812–2827.

[132] A. Donohoe, J. Marshall, D. Ferreira, and D. McGee. The Relationship between ITCZ Location and Cross-Equatorial Atmospheric Heat Transport : From the Seasonal Cycle to the Last Glacial Maximum. *J. Clim.*, 26, (2013), 3597–3618.

[133] R. S. Lindzen and S. Nigam. On the Role of Sea Surface Temperature Gradients in Forcing Low-Level Winds and Convergence in the Tropics. *J. Atmos. Sci.*, 44 (17), (1987), 2418–2436.

[134] K. A. Emanuel, J. D. Neelin, and C. S. Bretherton. On large-scale circulations in convecting atmospheres. *Quarter.y J. R. Meteorol. Soc.*, 120 (510), (1994), 1111–1143.

[135] M. Biasutti and A. Voigt. Seasonal and CO 2 -Induced Shifts of the ITCZ : Testing Energetic Controls in Idealized Simulations with Comprehensive Models. *J. Clim.*, 33, (2020), 2853–2870.

[136] S. P. Harrison, P. J. Bartlein, K Izumi, G Li, J. Annan, J Hargreaves, P Braconnot, and M Kageyama. Evaluation of CMIP5 palaeo-simulations to improve climate projections. *Nat. Clim. Chang.*, 5, (2015), 735–743.

[137] M. P. Byrne, A. G. Pendergrass, A. D. Rapp, and K. R. Wodzicki. Response of the Intertropical Convergence Zone to Climate Change : Location , Width , and Strength. *Curr. Clim. Chang. Reports*, 4, (2018), 355–370.

[138] T. Schneider, T. Bischoff, and G. H. Haug. Migrations and dynamics of the intertropical convergence zone. *Nature*, 513, (2014), 45–53.

[139] M. Biasutti et al. past , present and future monsoons. *Nat. Geosci.*, 11, (2018), 392–400.

[140] O. Adam, T. Bischoff, and T. Schneider. Seasonal and Interannual Variations of the Energy Flux Equator and ITCZ . Part I : Zonally Averaged ITCZ Position. *J. Clim.*, 29, (2016), 3219–3230.

[141] M. Wiedermann, J. F. Donges, J. Kurths, and R. V. Donner. Mapping and discrimination of networks in the complexity-entropy plane. *Phys. Rev. E*, 96 (4), (2017), 042304.

[142] S. M. Kang. Extratropical Influence on the Tropical Rainfall Distribution. *Curr. Clim. Chang. Reports*, 6, (2020), 24–36.

[143] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Sci. Rep.*, 2 (335), (2012), 1–9.

[144] D. M. J. Lazer et al. The science of fake news. *Science*, 359 (6380), (2018), 1094–1096.

[145] S. Myers and J. Leskovec. The Bursty Dynamics of the Twitter Information Network. *Proc. 23rd Int. Conf. World Wide Web*, (2014), 913–924. arXiv: `arXiv:1403.2732v1`.

[146] F. Baumann, P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini. Modeling echo chambers and polarization dynamics in social networks. *Phys. Rev. Lett.*, 124 (4), (2020), 048301. arXiv: `arXiv:1906.12325v2`.

[147] P. Lorenz-Spreen, B. M. Mønsted, P. Hövel, and S. Lehmann. Accelerating dynamics of collective attention. *Nat. Commun.*, 10 (1759), (2019).

[148] S. Fortunato. Community detection in graphs. *Phys. Rep.*, 486, (2010), 75–174. arXiv: `arXiv:0906.0612v2`.

[149] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466, (2010), 3–7.

[150] G. Palla, A.-L. Barabási, and T. Vicsek. Quantifying social group evolution. *Nature*, 446 (7136), (2007), 664–667.

[151] R. Cazabet, F. Amblard, and C. Hanachi. Detection of Overlapping Communities in Social Tagging Systems. *2010 IEEE Second Int. Conf. Soc. Comput.*, (2010), 309–314.

[152] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Tracking evolving communities in large linked networks. *PNAS*, 101 (1), (2004), 5249–5253.

[153] D. Greene, D. Doyle, and P. Cunningham. Tracking the Evolution of Communities in Dynamic Social Networks. *2010 Int. Conf. Adv. Soc. networks Anal. Min.*, (2010), 176–183.

[154] M. Rosvall, A. V. Esquivel, A. Lancichinetti, J. D. West, and R. Lambiotte. Memory in network flows and its effect on spreading dynamics and community detection. *Nat. Commun.*, 5 (4630), (2014).

[155] C. A. Yeung, N. Gibbins, and N. Shadbolt. Contextualising Tags in Collaborative Tagging Systems. *Proc. 20th ACM Conf. Hypertext Hypermedia*, (2009), 251–260.

[156] E. Ravacz and A.-L. Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67 (2), (2003), 026112.

[157] H. W. Kuhn. The Hungarian Method For The Assignment Problem. *Nav. Res Logist Q*, 2 (1-2), (1955), 83–97.

[158] S. Boulianne. Social media use and participation: a meta-analysis of current research. *Inf. Commun. Soc.*, 18 (5), (2015), 524–538.

[159] A. Acerbi. *Cultural evolution in the digital age*. Oxford University Press, 2019.

[160] M. Prensky. Digital natives, digital immigrants. *Horiz.*, 9 (5), (2001).

[161] T. Judd. The rise and fall (?) of the digital natives. *Australas. J. Educ. Technol.*, 34 (5), (2018), 99–119.

[162] B. A. Nardi, S. Whittaker, and E. Bradner. Interaction and outeraction: Instant Messaging in Action. *Proc. 2000 ACM Conf. Comput. Support. Coop. Work*, (2000), 79–88.

[163] D. Bouhnik and M. Deshen. WhatsApp Goes to School: Mobile Instant Messaging between Teachers and Students. *J. Inf. Technol. Educ. Res.*, 13, (2014), 217–231.

[164] G. Meikle. *Social media: Communication, sharing and visibility*. Routledge, 2016.

[165] J. H. Lipschultz. *Social media communication: Concepts, practices, data, law and ethics*. Taylor & Francis, 2017.

[166] J. M. Twenge, G. N. Martin, and B. H. Spitzberg. Trends in US Adolescents' media use, 1976–2016: The rise of digital media, the decline of TV, and the (near) demise of print. *Psychol. Pop. Media Cult.*, 8 (4), (2019), 329.

[167] Y. Liu, C. Kliman-Silver, and A. Mislove. The tweets they are a-Changin': Evolution of twitter users and behavior. *Proc. 8th Int. Conf. Weblogs Soc. Media, ICWSM 2014*, (2014), 305–314.

[168] B. Hutchins. The acceleration of media sport culture: Twitter, telepresence and online messaging. *Inf. Commun. Soc.*, 14 (2), (2011), 237–257.

[169] R. Schroeder. *Book Review: Social Theory after the Internet: Media, Technology and Globalization*. UCL Press, 2018.

[170] F. Bodendorf and C. Kaiser. Detecting opinion leaders and trends in online communities. *Proc. 2nd ACM Work. Soc. web search Min.*, (2009), 65–68.

[171] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the Dynamics of the News Cycle. *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, (2009), 497–506.

[172] Z. Tufekci. Engineering the public: Big data, surveillance and computational politics. *First Monday*, 19 (7), (2014).

[173] P. Lorenz-Spreen, S. Lewandowsky, C. R. Sunstein, and R. Hertwig. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nat. Hum. Behav.*, (2020).

[174] I. Waller and A. Anderson. Community embeddings reveal large-scale cultural organization of online platforms. *arxiv:2010.00590v2*, (2020). arXiv: 2010.00590.

[175] T. Alshaabi, D. R. Dewhurst, J. R. Minot, M. V. Arnold, J. L. Adams, C. M. Danforth, and P. S. Dodds. The growing echo chamber of social media: Measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020. *arxiv:2003.03667v4*, (2020). arXiv: 2003.03667.

[176] T. Yang, S. Majó-Vázquez, R. K. Nielsen, and S. González-Bailón. Exposure to news grows less fragmented with an increase in mobile access. *Proc. Natl. Acad. Sci.*, (2020), 202006089.

[177] K. Cheung and U. Ozturk. Synchronization of extreme rainfall during the Australian summer monsoon: Complex network perspectives. *Chaos*, 30 (063117), (2020).

[178] H. Su-Hong, F. Tai-Chen, G. Yan-Chun, H. Yan-Hua, W. Cheng-Guo, and G. Zhi-Quiang. Predicting extreme rainfall over eastern Asia by using complex networks. *Chin. Phys. B*, 23 (5), (2014), 059202.

[179] A. Agarwal, N. Marwan, R. Maheswaran, B. Merz, and J. Kurths. Quantifying the roles of single stations within homogeneous regions using complex network analysis. *J. Hydrol.*, 563, (2018), 802–810.

[180] U Ozturk, N Marwan, O Korup, H Saito, A Agarwal, M. J. Grossman, M Zaiki, and J Kurths. Complex networks for tracking extreme rainfall during typhoons. *Chaos*, 28 (075301), (2018).

[181] A. Agarwal, N. Marwan, U. Ozturk, and R. Maheswaran. *Unfolding Community Structure in Rainfall Network of Germany Using Complex Network-Based Approach*. Springer Singapore, 2019, pp. 179–193.

[182] J. Kurths, A. Agarwal, R. Shukla, N. Marwan, M. Rathinasamy, L. Caesar, R. Krishnan, and B. Merz. Unraveling the spatial diversity of Indian precipitation teleconnections via nonlinear multi-scale approach. *Nonlinear Process. Geophys. Discuss.*, 26, (2019), 251–266.

[183] A. Agarwal, N. Marwan, R. Maheswaran, U. Ozturk, J. Kurths, and B. Merz. Optimal design of hydrometric station networks based on complex network analysis. *Hydrol. Earth Syst. Sci.*, 24, (2020), 2235–2251.

Selbständigkeitserklärung

Ich erkläre, dass ich die Dissertation selbständig und nur unter Verwendung der von mir gemäß § 7 Abs. 3 der Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät, veröffentlicht im Amtlichen Mitteilungsblatt der Humboldt-Universität zu Berlin Nr. 42/2018 am 11.07.2018 angegebenen Hilfsmittel angefertigt habe.

Ich habe mich nicht anderwärts um einem Doktorgrad im Promotionsfach Physik beworben und besitze keinen Doktorgrad im Promotionsfach Physik.

Berlin, den 21. Mai 2021

Frederik Wolf